Corpus-based dialectometry: a methodological sketch

Benedikt Szmrecsanyi[1]

Abstract

In this paper, I introduce methodologies to tap corpora for exploring aggregate linguistic distances between dialects or varieties as a function of properties of geographic space. The paper describes the different steps necessary to obtain an appropriate corpus-based dataset (a so-called 'distance matrix'), and subsequently discusses several cartographic visualisation techniques – network maps, continuum maps and cluster maps – to project aggregate linguistic relationships to geography. In addition, the paper sketches some statistical methods to quantify these relationships. By way of example, a case study draws on the Freiburg Corpus of English Dialects – a major dialect corpus in which more than thirty traditional dialects of English from all over Great Britain are sampled. With a focus on regional variation in morphosyntax and on the basis of text frequencies of several dozen features, the study probes joint linguistic variability between the dialects sampled in the corpus.

1. Introduction: what is dialectometry?

Dialectometry is the branch of geolinguistics concerned with measuring, visualising and analysing aggregate dialect similarities or distances as a function of properties of geographic space. For seminal work, see Séguy (1971) – the paper that sparked the dialectometry enterprise; Goebl (1982, 1984, 2006) (the 'Salzburg School of Dialectometry'); and Nerbonne *et al*. (1999); Heeringa (2004); Nerbonne (2006) – the 'Groningen School of Dialectometry'. Whereas practitioners of traditional dialectology are dedicated to the study of 'interesting' – typically phonological or lexical – dialect phenomena, one feature at a time, and in a handful of

[1] Freiburg Institute for Advanced Studies (FRIAS), Albertstr. 19, 79104 Freiburg, Germany.
Correspondence to: Benedikt Szmrecsanyi, *e-mail*: bszm@frias.uni-freiburg.de

Corpora 2011 Vol. 6 (1): 45–76
DOI: 10.3366/corp.2011.0004

www.eupjournals.com/cor

dialects at most, dialectometrical inquiry endeavors to identify 'general, seemingly hidden structures from a larger amount of features' (Goebl and Schiltz, 1997: 13). This means that dialectometricians place a strong emphasis on quantification, cartographic visualisation and exploratory data analysis to infer patterns from feature aggregates. Empirically, the bulk of the dialectometrical literature draws on linguistic atlas material as its primary data source. For example, Goebl (1982) investigates joint variability in 696 linguistic features that are mapped in the *Sprach- und Sachatlas Italiens und der Südschweiz* (AIS), an atlas that covers Italy and southern Switzerland; Nerbonne *et al*. (1999) analyse aggregate pronunciational dialect distances between 104 Dutch and north Belgian dialects on the basis of 100 word transcriptions provided in the *Reeks Nederlands*(*ch*)*e Dialectatlassen* (RND). Some dialectometricians have also relied on dialect dictionaries (for example, Speelman and Geeraerts, 2008). Leinonen (2008), Heeringa *et al*. (2009) and Auer *et al*. (forthcoming) are rare examples of dialectometrical work which bases claims about aggregate accent differences on the analysis, auditory or acoustic, of actual speech samples. In any case, given that most dialect atlases and dictionaries focus on lexis and pronunciation, it should surprise nobody that much of the dialectometrical literature drawing on such material is biased towards lexis and pronunciation at the expense of morphological and, especially, syntactic variation (but see Spruit, 2005, 2006; and Spruit *et al*., 2009, for some recent atlas-based yet syntax-centred dialectometrical work).

There is no shortage of corpus-based research on regional variation in morphology and syntax. But, then again, while corpus-linguistic methodologies have increasingly found their way into the dialectological toolbox and while more and more dialect corpora are coming on-line (see Anderwald and Szmrecsanyi, 2009, for an overview), it is fair to say that the corpus-linguistic community is not exactly drowning in research that marries the qualitative-philological jeweller's-eye perspective inherent in the analysis of naturalistic corpus data with the quantitative-aggregational bird's-eye perspective that is the hallmark of dialectometrical research. In this study, I aim to remedy this shortcoming by discussing a methodology to conduct corpus-based dialectometry (see also Szmrecsanyi, 2008). As a case study to highlight the empirical potential of the methodology, we shall tap the Freiburg Corpus of English Dialects, a naturalistic speech corpus which contains samples of more than three dozen dialects from all over Great Britain. On the basis of this corpus, the study calculates a measure of aggregate dialectal distance that is based on joint variability of fifty-seven morphosyntactic dialect features. The investigation subsequently draws on a number of statistical analysis methods and utilises a range of cartographic projections to geography to aid interpretation of aggregate dialect distances. I wish to emphasise right at the outset, however, that this paper is of methodological talk, prioritising methodological aspects at the expense of detailed discussions and interpretations of results.

This paper has a distinct structure. Under Section 2, I present two arguments in favour of corpus-based dialectometry. Section 3 introduces the Freiburg Corpus of English Dialects. Section 4 discusses the design of an appropriate feature catalogue as the empirical basis for the aggregate analysis. Under Section 5, I discuss the feature extraction process and the creation of a so-called 'frequency matrix'. Section 6 addresses the actual aggregation process, which yields a so-called 'distance matrix'. I present ways to visually represent, analyse and interpret aggregate distances and similarities between dialects under Section 7. Section 7.1 offers some concluding remarks.

2. Why corpus-based dialectometry?

The marriage of corpus-based variationist research and aggregative-dialectometrical analysis techniques is desirable for two principal reasons.

First, multidimensional objects, such as dialects, call for aggregate analysis techniques. So-called 'single-feature-based studies' (Nerbonne, 2009: 176), with their atomistic focus on, typically, just one feature, are fine when it is the features themselves that are of analytic interest. They are woefully inadequate, however, when it comes to characterising multidimensional objects such as dialects or varieties (or relations between them). Outside linguistics, this sort of inadequacy is well-known: taxonomists, for instance, typically categorise species not on the basis of a single morphological or genetic criterion, but on the basis of many; economists assess the economic climate not on the basis of individual macroeconomic indicators (e.g., unemployment), but also on inflation, GDP per capita, interest rates, and so on. The problem with single-feature-based studies – in linguistics as well as everywhere else – is that feature selection is ultimately arbitrary (see Viereck, 1985: 94), and that the next feature down the road may or may not contradict the characterisation suggested by the previous feature. Thus, there is no guarantee that different dialects will exhibit the same distributional behavior with regard to different features; isoglosses do not necessarily overlap (see Bloomfield, 1984 [1933]: 329). In addition, individual features may have fairly specific quirks to them that are irrelevant to the big picture. This is why 'single-feature studies risk being overwhelmed by noise, i.e., missing data, exceptions, and conflicting tendencies' (Nerbonne, 2009: 193). So, the aggregate perspective – in Goebl's (2006: 415) parlance, 'the synthetic interpretation' of linguistic data – is called for when the analyst's attention is turned to the forest, not the trees. Aggregation mitigates the problem of feature-specific quirks, irrelevant statistical noise and the problem of inherently subjective feature selection, and, thus, provides a more robust linguistic signal.

Second, compared to dialect atlas material (and I subsume, here, dialect dictionary material), corpora yield a more realistic linguistic signal.

Atlas-based dialectometry typically aggregates observations such as 'in the Yorkshire dialect, the lexeme *bus* is typically pronounced /bʊs/', while corpus-based (that is to say, frequency-based) approaches seek generalisations along the lines of 'in Nottinghamshire English, multiple negation is twice as frequent (six occurrences per ten thousand words) in actual speech than in Yorkshire English (three occurrences per ten thousand words)'. The atlas-based method has undeniable advantages. We emphasise, in particular, a fairly widespread availability of data sources and superb areal coverage. By contrast, dialect corpora are a rarer species, and their areal coverage is typically inferior to dialect atlases. Having said that, as a data source, corpora appear to have two major advantages over dialect atlases. First and foremost, the atlas signal is categorical, exhibits a high level of data reduction, and may thus be less accurate than the corpus signal, which can provide graded frequency information (see Wälchli, 2009; and Holman *et al*., 2007: 413). This highlights the most crucial difference between atlas-based and corpus-based dialectometry: corpus-based dialectometry is frequency-based dialectometry in its purest form (which is why the approach outlined in this paper bears a certain similarity to the method of Hoppenbrouwers and Hoppenbrouwers, 2001, discussed in some length in Heeringa, 2004: 16–20). The point is that although the exact cognitive status of text frequencies is admittedly still unclear (for example, we do not currently know about the precise extent to which corpus frequencies correlate with psychological entrenchment; see Arppe *et al*., 2010), we do claim that text frequencies match better with perceptual salience than discrete atlas classifications; this is true even though some varieties of atlas-based dialectometry derive – with considerable computational effort – some form of commonness weighting (for instance, at the phonetic segment level) from the atlas signal. Second, we note that the atlas signal is non-naturalistic and, basically, meta-linguistic in nature. It typically relies on elicitation and questionnaires, and is analytically twice removed (through fieldworkers *and* atlas compilers) from the analyst. By contrast, text corpora provide more direct access to language form and function, and may thus yield a more realistic and trustworthy picture (see Chafe, 1992: 84; and Leech *et al*., 1994: 58). The well-known major intrinsic drawback of the corpus-based method is that it is unable to deal with rare phenomena (see Haspelmath, 2009: 157–58; and Penke and Rosenbach, 2007: 489) – but, then again, it is arguable whether phenomena that are so infrequent that they cannot be described on the basis of a major text corpus should have a place in an aggregate analysis at all.

3. The data source: the Freiburg Corpus of English Dialects, FRED

By way of a sample analysis, this paper will tap the Freiburg Corpus of English Dialects (FRED; see Hernández, 2006; and Szmrecsanyi and Hernández, 2007, for details). The version of the corpus used in this study

contains 368 individual texts and spans approximately 2.44 million words of running text, consisting of samples (mainly transcribed so-called 'oral history' material) of dialectal speech from a variety of sources. Most of these samples were recorded between 1970 and 1990; in the majority of cases, a fieldworker interviewed an informant about life, work, *etc*., in the past. The 431 informants sampled in the corpus are typically elderly people with a working-class background – so-called 'NORMs' (non-mobile old rural males, see Chambers and Trudgill, 1998: 29). The interviews were conducted in 156 different locations (that is, villages and towns) in thirty-four different pre-1974 counties in Great Britain, including the Isle of Man and the Hebrides. The level of areal granularity investigated in this study will be at the level of the county. This leaves us with thirty-four objects (i.e., dialects or measuring points, which are listed in Table 1) that will be exemplarily subjected to dialectometrical analysis in the subsequent sections. Note that the corpus is annotated with longitude and latitude information for each of the locations that were sampled. From this annotation, county coordinates (mean longitude and latitude) can be calculated by computing the arithmetic mean of all the location coordinates associated with a particular county.

4. The empirical foundation: defining the feature catalogue

The first step towards dialectometrical analysis is defining the feature catalogue as the empirical basis for the corpus-cum-aggregation endeavor. In keeping true to the spirit of dialectometrical analysis, the goal is to base the analysis on as many features as possible. In the case study at hand, I surveyed the dialectological, variationist and corpus-linguistic literature on morphosyntactic variability in varieties of English, and identified suitable dialect phenomena. This resulted in a list of $p = 57$ features, which overlaps with, but is not identical to, the comparative morphosyntax survey in Kortmann and Szmrecsanyi (2004) and the battery of morphosyntax features covered in the Survey of English Dialects (see Orton *et al*., 1978; and Viereck *et al*., 1991). The features in the catalogue fall into eleven major grammatical domains:

(*i*) pronouns and determiners (e.g., non-standard reflexives);
(*ii*) the noun phrase (e.g., zero plural endings);
(*iii*) primary verbs (e.g., the verb TO DO);
(*iv*) tense and aspect (e.g., the present perfect with auxiliary BE);
(*v*) modality (e.g., epistemic/deontic *must*);
(*vi*) verb morphology (e.g., non-standard weak past tense and past participle forms);
(*vii*) negation (e.g., *never* as a preverbal past tense negator);
(*viii*) agreement (e.g., non-standard WAS);
(*ix*) relativisation (e.g., the relative particle *what*);

Map label	County	*A-priori* dialect area	Mean longitude	Mean latitude	Number of words sampled in FRED
ANS	Angus	Sc Lowlands	–2.627	56.659	19,933
BAN	Banffshire	Sc Lowlands	–2.949	57.543	5,671
CON	Cornwall	southwest of E.	–5.502	50.175	107,290
DEN	Denbighshire	Wales	–3.743	53.146	5,789
DEV	Devon	southwest of E.	–3.681	50.378	97,229
DFS	Dumfriesshire	Sc Lowlands	–3.839	55.003	10,019
DUR	Durham	north of E.	–1.703	54.89	28,086
ELN	East Lothian	Sc Lowlands	–2.954	55.945	40,193
GLA	Glamorganshire	Wales	–3.634	51.641	53,229
HEB	Hebrides	Hebrides	–7.038	57.502	73,209
MAN	Isle of Man	Isle of Man	–4.446	54.257	10,945
KCD	Kincardineshire	Sc Lowlands	–2.465	56.974	7,514
KEN	Kent	southeast of E.	0.835	51.246	177,055
LAN	Lancashire	north of E.	–2.73	53.653	205,475
LEI	Leicestershire	English Midlands	–1.623	52.752	5,864
LND	London	southeast of E.	–0.068	51.504	110,878
MDX	Middlesex	southeast of E.	–0.382	51.594	31,794

Table 1: $n = 34$ objects (i.e., FRED counties/dialects) considered in the present study: map labels, membership in *a-priori* dialect areas roughly following Trudgill's dialect division on pronunciational grounds (Trudgill, 1999: Map 9), mean longitude, mean latitude, textual coverage (running words) in FRED

Map label	County	*A-priori* dialect area	Mean longitude	Mean latitude	Number of words sampled in FRED
MLN	Midlothian	Sc Lowlands	–3.265	55.918	32,040
NBL	Northumberland	north of E.	–1.68	55.302	30,771
NTT	Nottinghamshire	English Midlands	–1.055	53.011	150,889
OXF	Oxfordshire	southwest of E.	–1.598	51.787	15,139
PEE	Peebleshire	Sc Lowlands	–3.377	55.721	14,975
PER	Perthshire	Sc Lowlands	–3.53	56.368	20,960
ROC	Ross and Cromarty	Sc Highlands	–4.776	57.808	10,495
SAL	Shropshire	English Midlands	–2.471	52.653	169,133
SEL	Selkirkshire	Sc Lowlands	–3.002	55.502	9,365
SFK	Suffolk	southeast of E.	1.699	52.555	312,600
SOM	Somerset	southwest of E.	–2.792	51.112	208,264
SUT	Sutherland	Sc Highlands	–4.676	58.144	11,025
WAR	Warwickshire	English Midlands	–1.968	52.574	8,271
WES	Westmorland	north of E.	–2.962	54.428	157,590
WIL	Wiltshire	southwest of E.	–2.031	51.259	186,239
WLN	West Lothian	Sc Lowlands	–3.784	56.001	18,418
YKS	Yorkshire	north of E.	–1.174	54.424	90,963

Table 1 (*continued*): $n = 34$ objects (i.e., FRED counties/dialects)

(*x*) complementation (e.g., unsplit *for to*); and,
(*xi*) word order and discourse phenomena (e.g., lack of auxiliaries in *yes*/*no* questions).

A detailed discussion of the features in the catalogue is beyond the scope of this paper, but Appendix A provides the complete list of features.

A few comments on the case study's criteria for feature selection are needed, however. For a feature to be included in the catalogue, it did not matter whether the feature had previously been reported as having geographic variation or not. For instance, Feature 31 (the negative suffix, –*nae*) has a very clear and well-known regional distribution, but Feature 10 (preposition stranding) does not, according to the literature. Also note that the catalogue contains fairly definite, and thus somewhat salient, non-standard features, which tend to be either largely present or absent – Feature 31 (the negative suffix, –*nae*) is again a good example – but also encompasses features whose variation is more statistical in nature, and thus arguably less salient (for example, Features 8 and 9 on genitive variation). The features included in the catalogue also differ in terms of their 'standardness': Feature 2 (standard reflexives), for instance, is examined with respect to the text frequency of perfect standard forms, while Feature 28 (non-standard weak verb forms) is not really acceptable in Standard English. In short, the feature catalogue seeks to span as many features as possible, regardless of their geographic distribution, the scope of their variability and their standardness. The rationale is that non-geographic and/or random variability will cancel out in the aggregate view. For practical purposes, however, two criteria had to be met for a candidate feature to be included in the catalogue. First, to ensure statistical robustness of text frequencies, the feature had to be relatively frequent. Specifically, the feature had to have a raw frequency of at least 100 raw hits in FRED. This criterion ruled out demonstrably infrequent phenomena such as resumptive relative pronouns, double modals, the relativiser *as*, and so on. Second, a candidate feature also had to be extractable – subject to a reasonable input of labour resources – by a human coder. This is why, for example, many hard-to-retrieve null phenomena (such as zero relativisation) or features where semantics must be taken into consideration (such as gendered pronouns) are not included in the catalogue.

5. Data mining: extracting feature frequencies and creating a frequency matrix

The second step consists of extracting feature frequencies and creating a frequency matrix. In terms of this study's feature catalogue, thirty-one sufficiently 'surfacy' features (e.g., the negator *ain*'*t*) were extracted right away by software; and twenty-six features in the catalogue (e.g., *don*'*t* with third person singular subjects) required more or less substantial

manual disambiguation prior to extraction. Szmrecsanyi (2010) provides detailed coding schemes and discusses the technicalities of the extraction process for all fifty-seven features in the catalogue. Once feature frequencies are extracted, the analyst will normalise text frequencies – for example, to frequency per 10,000 words – if, as is the case with most relevant corpora, textual coverage of individual dialects varies. At this stage, we also recommend a *log*-transformation as a customary method to de-emphasise large frequency differentials and to alleviate the effect of frequency outliers (see Shackleton, 2007: 43), thus increasing the reliability of the frequency matrix. To illustrate, in FRED, the county of Cornwall has a textual coverage of twelve interviews totalling about 107,000 words of running text (excluding interviewer utterances). In this material, Feature 34 (negative contraction, e.g., *they won't do anything*) occurs 326 times, which translates into a normalised text frequency of $326 \times 10{,}000/107{,}000 \approx 30$ occurrences per ten thousand words. A *log*-transformation of this frequency yields a value of $log_{10}(30) \approx 1.5$.[2] This is the figure that characterises this specific measuring point (Cornwall) with regard to Feature 34.

The next step is to create an $N \times p$ frequency matrix in which the N objects (that is, dialects) are arranged in rows and the p features in columns, such that each cell in the matrix specifies a particular (normalised and *log*-transformed) feature frequency. Our case study thus yields a 34×57 frequency matrix: thirty-four British English dialects, each characterised by a vector of fifty-seven text frequencies.

At this point, the analyst must assess the *reliability* of the frequency matrix: are the features included in the catalogue a heterogeneous 'mixed bag' (for which an aggregate analysis would be meaningless), or is there a sufficient degree of consistency? Calculating a statistic known as *Cronbach's* α (see Cronbach, 1951; and Nunnally, 1978) can address this issue.[3] Cronbach's α is, technically speaking, a coefficient measuring the average inter-item (in our case, inter-feature) correlation. Cronbach's α can take values between negative infinity and 1. An α value of 0 indicates that the features under investigation are not at all related, and a value of 1 means that all the features are perfectly correlated. Higher α values thus indicate a higher rate of reliability of the frequency matrix. By convention (in dialectometry (see Heeringa, 2004: 173) and elsewhere (see, for example, Bland and Altman, 1997)), researchers aim for Cronbach's α values of .7 or higher. If a given frequency matrix yields a Cronbach's α value smaller than .7, there is a problem that should be addressed by expanding or altering the composition of the feature catalogue. My case study's 34×57 frequency matrix yields a Cronbach's α value of .77, a score that comfortably passes the conventional threshold.

---

[2] Text frequencies of 0 (for instance, Feature 4 (archaic *ye*, as in *ye'd dancing every week*) does not occur in material from Cornwall) were rendered as 0.1, which yields a *log*-transformed value of $log_{10}(0.1) = -1$.

[3] Standard statistical software packages, such as R and SPSS, can easily calculate the Cronbach's α statistic.

6. Aggregation: obtaining a distance matrix

The task before us now is to convert the $N \times p$ frequency matrix into an $N \times N$ distance matrix. This transformation is an aggregation step, in that the resulting distance matrix abstracts away from individual feature frequencies and specifies pairwise distances between the objects considered (similar to distance tables to be found in, for example, road atlases). How do we calculate aggregate distances? Standard software packages offer a bewildering array of distance measures. Yet given the continuous nature of corpus-derived frequency vectors, I advocate use of the well-known and fairly straightforward Euclidean Distance Measure (see, for instance, Aldenderfer and Blashfield, 1984: 25) unless there is a good reason not to use it. Drawing on the Pythagorean Theorem (see Nishisato, 2007: 77), the Euclidean Distance Measure defines the distance between two objects, a and b, as the square root of the sum of all p squared frequency differentials:

$$d(a,b) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_p - b_p)^2} = \sqrt{\sum_{i=1}^{p} (a_i - b_i)^2}$$

Here, p is the number of features, a_1 is the frequency of feature 1 in object a, b_1 is the frequency of feature 1 in object b, a_2 is the frequency of feature 2 in object a, and so on. The Euclidean Distance Measure is, for one thing, interpretationally convenient: in two-dimensional space, it yields the distance between two points that one would measure with a ruler, which is why the measure is also sometimes referred to as 'ruler distance' (Giles, 2002: 139); furthermore, the Euclidean Distance Measure is theory-neutral in that all features receive the same weight in the distance calculation. Having said that, I would like to stress that bigger frequency differentials receive proportionally more weight than smaller frequency differentials, which must appear as a desirable property to all those who believe that corpus frequencies mirror some sort of psychological and perceptual reality.

The chart in Figure 1 illustrates the aggregation process. In step (1), we begin with a fictional 3×2 frequency matrix, which contains six cells specifying frequencies of two features in three dialects. In step (2), we calculate three distances: the distance between dialects a and b (which we commonsensically define as identical to the distance between dialects b and a), the distance between dialects a and c, and the distance between dialects b and c. In step (3), we enter these distances into a 3×3 distance matrix, which has $3 \times (3–1)/2 = 3$ unique cells, (i.e., dialect/dialect pairings). The other cells are redundant in that the distance between a given dialect and itself is always zero, and the distances in the upper right half of the matrix would mirror the distances in the lower left half of the matrix.

① **the frequency matrix**

	text frequencies feature 1	text frequencies feature 2
dialect *a*	11	8
dialect *b*	5	2
dialect *c*	1	7

↓

② **aggregation via the Euclidean distance measure**

$$d(a,b) = \sqrt{(11-5)^2+(8-2)^2} = 8.5$$
$$d(a,c) = \sqrt{(11-1)^2+(8-7)^2} = 10.0$$
$$d(b,c) = \sqrt{(5-1)^2+(2-7)^2} = 6.4$$

↓

③ **the distance matrix**

	dialect *a*	dialect *b*	dialect *c*
dialect *a*			
dialect *b*	8.5		
dialect *c*	10.0	6.4	

Figure 1: Converting a fictional 3×2 frequency matrix into a 3×3 distance matrix utilising Euclidean distance as an aggregation measure

7. Visualisation, analysis and interpretation

Distance matrices can be analysed in a myriad of ways, not all of which may make sense for a particular set of research questions. This section sketches some ways to represent visually, to analyse and to interpret distance matrices (geo)linguistically. Under Section 7.1, I utilise so-called 'network maps' to project aggregate dialect distances and similarities to geography. In Section 7.2, I rely on 'continuum maps' to probe the extent to which joint dialectal variability is structured in terms of a dialect continuum. Section 7.3 draws on 'cluster maps' to explore the existence of dialect areas. Section 7.4 marshals correlative statistical analysis techniques to gauge the explanatory power of a number of language-external distance measures.[4]

[4] On a technical note, all cartographic projections and most of the non-trivial statistical analyses (such as multidimensional scaling and cluster analysis) presented in this section were created using Peter Kleiweg's *RuG/L04* dialectometry software package (available for free at http://www.let.rug.nl/∼kleiweg/L04/). The input required by *RuG/L04* is (*i*) the longitude/latitude coordinates provided in Table 1, (*ii*) a (linguistic) distance matrix, and (*iii*) a polygon map – which can be created using Google Earth – that defines the boundaries of a land mass and/or political borders. Note, along these lines, that there is another major dialectometry package: The *Visual DialectoMetry* (VDM) software developed in Salzburg (Haimerl, 2006). VDM is also free and comes, as an added bonus, with a graphical user interface.

measuring point pairings	561
mean	5.41
standard deviation	1.11
minimum	2.32
median	5.40
maximum	8.14
skewness	−.06
kurtosis	−.37

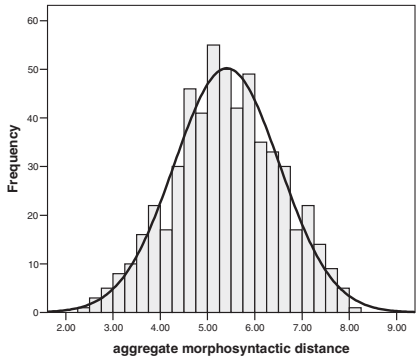

Figure 2: Aggregate morphosyntactic distances: summary statistics (left) and histogram (right)

7.1 Projecting aggregate distances and similarities to geography

In this section, I will discuss how to investigate the distribution of aggregate dialect distances. In this spirit, the table under Figure 2 provides a number of summary statistics which describe the distribution of morphosyntactic dialect distances in Great Britain. Our distance matrix spanning $n = 34$ FRED dialects yields $34 \times 33/2 = 561$ pairwise distances. Mean morphosyntactic distance is 5.41 Euclidean Distance points. This distance roughly corresponds to the distance between two hypothetical dialects, *a* and *b*, where dialect *a* attests a normalised text frequency of two hits per 10,000 words for each of the fifty-seven features, while dialect *b* attests a normalised text frequency of approximately ten hits per 10,000 words for each of the fifty-seven features. As for the dataset-internal dispersion around the mean, we are dealing with a standard deviation of 1.11. Given that the distances are, as we shall see shortly, normally distributed, this is another way of saying that roughly two thirds of the 561 county/county pairings score a distance within 1.11 points of the mean, and that 95 percent of all pairwise distances do not deviate more than 2.22 points from the mean. The minimum observable distance in the dataset is 2.32 points (this happens to be the morphosyntactic distance between the dialects spoken in the county of Somerset and the county of Wiltshire, two neighbouring counties located in the southwest of England). The median distance is 5.40 Euclidean Distance points. This is the distance that separates the higher half of the distance sample from the lower half. The maximum observable distance in the dataset is 8.14 points, which is the distance between the dialects spoken in the county of Denbighshire in Wales and the county of Kincardineshire in the Scottish Lowlands.

Are pairwise morphosyntactic distances between the counties normally distributed? The histogram under Figure 2, which plots the frequency of a number of distance brackets, suggests that they roughly are.

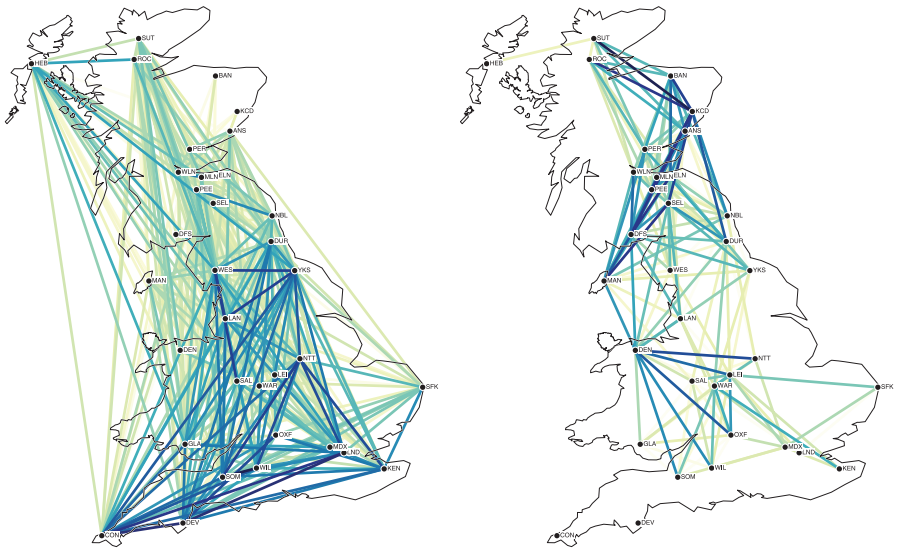

Figure 3: Projecting aggregate morphosyntactic dialect relationships to geography: network maps. Link blueness is directly proportional to dialectal similarity (left) or dialectal distance (right)

Numerically speaking, the skewness value of –.06 suggests that there is only a very slight negative skew, such that there is a greater number of larger distances than smaller distances. As for 'peakedness', the kurtosis value of –.37 indicates that the distribution of distances is a bit flatter than it would be in a perfectly normal distribution. Having said that, skewness and kurtosis values of ±1.0 are, by convention (see Meyers *et al*., 2006: 90), taken to be indicative of a normal – albeit not perfectly normal – distribution.

The maps in Figure 3 are Groningen-style Network Maps (see Nerbonne and Heeringa, 1997) that project dialect distances to geography without much statistical ado.[5] The simple idea behind the left map in Figure 3 is that dialects that are close linguistically are linked by darker, more blueish lines, while linguistically more distant dialects are linked by proportionally lighter, more yellowish lines. Visual inspection of the map reveals that we are dealing with a network of comparatively strong morphosyntactic links in England, and with a somewhat looser network structure in Scotland. Within England, we observe particular strong link bundles in the south and in the north. Northumberland seems to link well to some Scottish measuring points, and, turning back to the literature, we note that both Ellis (1889) and Trudgill (1999) actually regard the traditional dialect spoken in northern Northumberland as a Scots variety. The Hebrides have strong morphosyntactic ties to measuring points all over Great Britain;

[5] The maps were created using the *R*u*G*/*L04* package's *maplink* module.

notice in this connection that as a Scottish Highlands variety, Hebridean English is a relatively young dialect which Trudgill, for example, does not in fact categorise as a traditional dialect on account of the fact that it has 'become English-speaking only relatively recently' (Trudgill, 1999: 5). It is not particularly surprising, therefore, that Hebridean English lacks a clear-cut morphosyntactic profile of its own and bears similarities to dialects all over the place.

As a mirror image of sorts to the previous discussion, the right map in Figure 3 (a reverse network map) highlights morphosyntactic dissimilarities in the dataset. This particular map omits links between dialects that are more than 250 km apart, which mainly serves presentational purposes by enhancing readability without swamping the reader with an abundance of dissimilarity links. In Figure 3's reverse network map we observe, first, the striking tangle of dissimilarities covering much of northern England and Scotland. In this connection, it is especially Banffshire and Kincardineshire that radiate strong beams of dissimilarity to other measuring points in the FRED network. In England, there is a web of modest dissimilarities involving measuring points in the Midlands (Shropshire, Warwickshire, Leicestershire and Nottinghamshire). Warwickshire is, in addition, dissimilar to a number of sites in the southwest of England (Somerset, Wiltshire and Oxfordshire). As for Wales, observe the strong signal of dissimilarity emanating from the county of Denbighshire in northern Wales; by contrast, Glamorganshire in southern Wales is fairly inconspicuous. This difference between the two measuring points in Wales is not surprising: we know that there are robust lexical, phonological and grammatical differences between dialects in the north and in the south of Wales (Penhallurick, 1993), many of which can be traced back to an influx of southwestern English speakers to the south of Wales beginning as early as the end of the eleventh century AD (Penhallurick, 2004: 98). By comparison, Denbighshire English in the north of Wales is a younger dialect with less well-established historical links to English English dialects.

In this section, I have provided a first impression of the overall picture by considering summary statistics as well as network maps to represent aggregate distances in geographic map space. In what follows, I push deeper into the geographic structure of linguistic variation, subjecting the distance matrix to a good deal of statistical processing and subsequently projecting the output to geography.

7.2 Exploring dialect continua

Many dialectologists and geolinguists assume that geographic proximity predicts dialectal similarity. Nerbonne and Kleiweg (2007: 154) refer to this axiom as the Fundamental Dialectology Principle. In this section, I present ways to depict the extent to which linguistic distance is directly proportional

to geographic distance such that there are 'no real boundaries, but only gradual transitions' (Bloomfield, 1984 [1933]: 341).

To approach this issue cartographically, I turn to so-called Continuum Maps, a signature visualisation technique developed in Groningen (see Nerbonne *et al*., 1999; and Heeringa, 2004). On the cartographic side, I set the scene by utilising customary Voronoi tessellation (Voronoi, 1907; and Goebl, 1984) to assign each dialect site on the map a convex polygon such that each point within the polygon is closer to the generating dialect site than to any other dialect site. Notice that when areal coverage is very fine-grained (as is usually the case in dialect atlases), it makes sense to tessellate map space into Voronoi polygons exhaustively. However, my case study covers Great Britain with $n = 34$ sampling points, which is why I prefer to limit the radius of the Voronoi polygons to approximately 50 km in order to do visual justice to the areal coverage of the dialect corpus. The next step is a computational one and subjects the data to Multidimensional Scaling (MDS, see Embleton, 1993; and Kruskal and Wish, 1978). MDS is an exploratory statistical technique used to reduce a higher-dimensional dataset to a lower-dimensional representation which is more amenable to visualisation. The task here is to scale down a $N-1$ dimensional distance matrix (in which each object is characterised by its distance to the other $n-1$ objects in the matrix) to a three-dimensional representation, in which each object has a coordinate in three artificial MDS dimensions. These coordinates are then mapped to the red–green–blue colour scheme, giving each of the Voronoi polygons a distinct hue. On the interpretational plane, then, smooth colour transitions between dialect polygons emphasise the continuum-like nature of the dialect landscape; abrupt colour transitions point to the necessity of alternative explanations.[6]

Turning back to our case study, in Figure 4 we find two continuum maps that explore and, in fact, correlate (see Goebl, 2005) the dialect landscape to the geographic landscape in Great Britain. The left-hand map is based on scaling a distance matrix and details not linguistic distances but as-the-crow-flies geographic distances between dialect sites, and thus depicts, for reference purposes, a perfect continuum. The right-hand map in Figure 4 visually represents an MDS solution that scales actual morphosyntactic distances. Statistically speaking, the MDS distances underlying the left visualisation capture 100 percent ($r = 1.0$) of the variance in the original as-the-crow-flies distance matrix. The MDS solution depicted in the right-hand map captures about 89.5 percent ($r = .95$) of the distance variance in the original linguistic dataset, which is a fairly good score.

In all, the mosaic pattern in the morphosyntax continuum map suggests that the morphosyntactic dialect landscape in Great Britain is less continuum-like than it could be. It is true that there are some fairly nice micro-continua, especially so in the southwest of England and in the central

[6] The visualisation and statistical analysis techniques in this section draw on the *R*u*G*/*L04* package's *mds* (method: Kruskal's MDS) and *maprgb* modules.

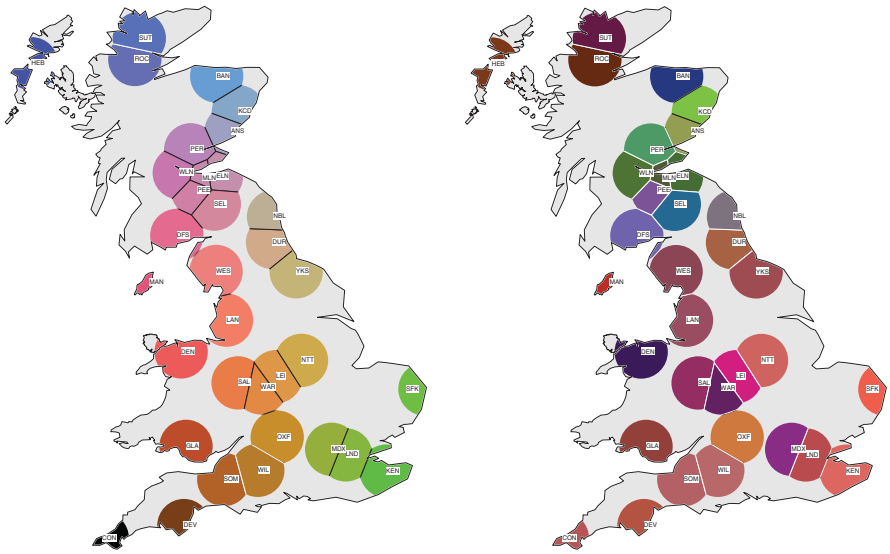

Figure 4: Continuum maps: a perfect continuum for reference purposes (left) *versus* the actual linguistic situation (right). Similar colour hues indicate morphosyntactic similarity. Input left projection: as-the-crow-flies distances (correlation with distances in the original distance matrix: $r = 1.00$). Input right projection: morphosyntactic distances (correlation with distances in the original distance matrix: $r = .95$)

and northern Scottish Lowlands. Note also that dialects spoken in the north of England fade rather smoothly into southern Scottish Lowlands dialects. But we also observe rather abrupt transitions between the central Scottish Lowlands – comprising dialects spoken in West Lothian, Midlothian and East Lothian – and southern Scottish dialects (Peebleshire and Selkirkshire). In England, the dialects spoken in Middlesex and Warwickshire are outliers. In Wales, it is Denbighshire that does not fit into the picture.

At this point, it is instructive to abandon the aggregate perspective for a moment and to re-consider the actual features on which the analysis is based. In this spirit, to aid interpretation of continuum maps the analyst can correlate frequency vectors with MDS dimensions to identify those features that are most robustly implicated in the overall dimensionality (see Heeringa, 2004: 266–71). In this case study, this produces the following top correlations between colour shades and feature frequencies in the right-hand map displayed in Figure 4:

- Increased text frequencies of Feature 33 (multiple negation, as in you *didn't want no beer*) correlate best with MDS dimension 1, which yields reddish tones ($r = .82$, $p < .001$).

- Increased text frequencies of Feature 31 (the negative suffix *–nae*, as in *I cannae mind of ever being laid off*) correlate best with MDS dimension 2, which yields greenish tones ($r = .75$, $p < .001$).
- Increased text frequencies of Feature 28 (non-standard weak past tense and past participle forms, as in *we runned up a bill*) correlate best with MDS dimension 3, which yields blueish tones ($r = .72$, $p < .001$).

By way of an interim summary, we have seen in this section that viewing aggregate morphosyntactic variability in a dialect continuum perspective can be instructive. Yet this approach does not necessarily tell the whole story: the existence of abrupt dialect transitions, as in Figure 4, suggests that linguistic variability may be organised in terms of dialect areas rather than continua. The next section is dedicated to investigating this hypothesis more closely.

7.3 The dialect area scenario

The tacit assumption guiding the foregoing discussion was that linguistic similarity between dialects is inversely proportional to geographic distance between dialects. There is, however, an alternative view, according to which dialect landscapes may be geographically organised along the lines of geographically coherent and linguistically homogeneous 'areas within which similar varieties are spoken' (Heeringa and Nerbonne, 2001: 375). In this view, we should find linguistic boundaries between, rather than within, dialect areas. In this section, we discuss methods to explore this view.

The dialect area scenario lends itself to visualisation using Cluster Maps, a cartographic technique that is common in all strands of dialectometry and which projects the outcome of cluster analysis to geography (see, for example, Goebl, 2007: Map 18; and Heeringa, 2004: Figure 9.6). Exactly as with continuum maps, the starting point is a Voronoi tesselation of map space. Subsequently, the $N \times N$ distance matrix is subjected not to MDS, but to Hierarchical Agglomerative Cluster Analysis (see Jain *et al*., 1999), a statistical technique that is used to group a number of objects (in this study, dialects) into a smaller number of discrete clusters. While there are many different clustering algorithms, I prefer Ward's Minimum Variance Method (Ward, 1963), an algorithm that tends to create small and even-sized clusters and which is popular both in corpus linguistics (for example, Gries and Wulff, 2005) and in dialectometry (see, for instance, Goebl, 2008).[7] Cluster analysis initially yields a so-called dendrogram (see Figures 5 and 6), which

[7] Simple clustering can be unstable, so the analysis in this section utilises a procedure known as 'clustering with noise' (Nerbonne *et al*., 2008): the original distance matrix is clustered repeatedly, adding some random amount of noise ($c = \sigma/2$) in each run. This exercise yields a cophenetic distance matrix which provides consensus (and thus more stable) cophenetic distances between dialects.

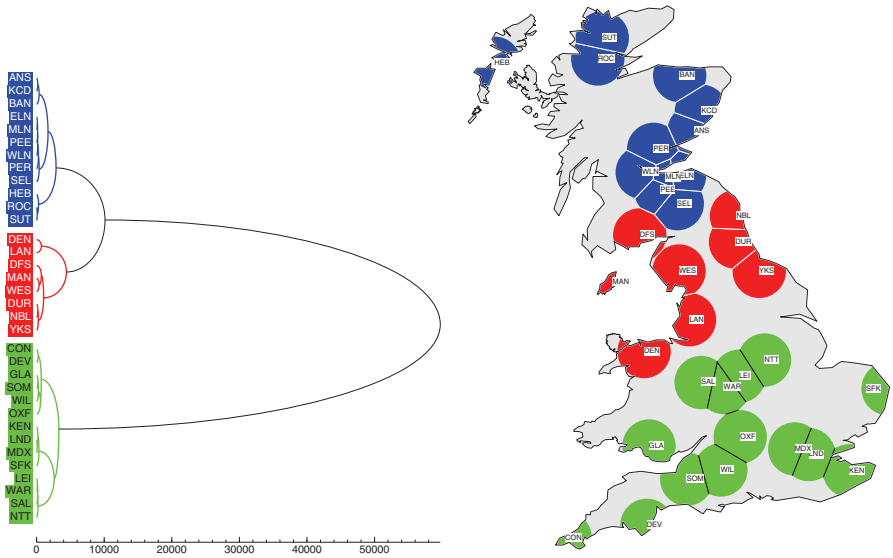

Figure 5: A perfect regionalisation for reference purposes: clustering as-the-crow-flies distances (hierarchical agglomerative cluster algorithm: WARD). Displayed: three-cluster solution. Left: dendrogram. Right: cluster map. Colours indicate regional membership

depicts cophenetic distances between the clustered objects. The optimal number of clusters is determined by, for example, diagramming the number of clusters against the fusion coefficient and spotting the 'elbow' in the resulting graph (see Aldenderfer and Blashfield, 1984: 54). Finally, each of the clusters is assigned a distinct colour and the Voronoi polygons are colourised accordingly.[8]

Applying these steps to the dataset on dialect variability in Great Britain yields Figures 5 and 6. Figure 5, which is based not on morphosyntactic but on as-the-crow-flies geographic distances, will serve as the non-linguistic reference point for my discussion. The map suggests that on strictly geographic grounds and according to Ward's method, Great Britain can be partitioned into three coherent areas: a green region comprising the south of England plus the county of Glamorganshire in southern Wales; a red region containing the north of England plus the county of Denbighshire in northern Wales plus the county of Dumfriesshire in southern Scotland; and a blue region encompassing Scotland minus the county of Dumfriesshire.

[8] The visualisation and statistical analysis techniques in this section draw on the *RuG/L04* package's *cluster* module (which implements clustering with noise; cf. footnote 7), as well as on the *mapclust* and *den* modules.

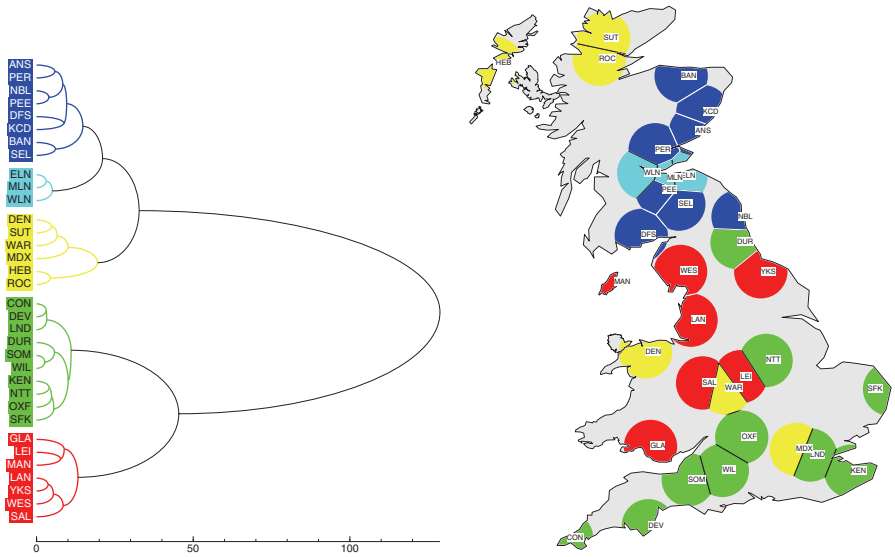

Figure 6: Actual morphosyntax clusters: Clustering morphosyntactic distances (hierarchical agglomerative cluster algorithm: WARD). Displayed: 5-cluster solution. Left: dendrogram. Right: cluster map. Colours indicate dialect area or dialect grouping membership

Compare this landscape to Figure 6, which visually depicts a five-cluster regionalisation on morphosyntactic grounds. There is clearly some similarity between the geographic and linguistic partitioning, although I note that there is also a good deal of geographic incoherence in the morphosyntax division. A more detailed account would highlight the following differences between the maps in Figures 5 and 6:

- The yellow dialect grouping in the morphosyntax map encompasses some isolated outliers in England (Middlesex and Warwickshire), Wales (Denbighshire) and a geographically coherent sub-cluster of Scottish Highland dialects (Ross and Cromarty, and Sutherland) plus the Hebrides.
- The morphosyntax visualisation has a small yet geographically coherent central Scottish Lowlands dialect area (in light blue), comprising the counties of East Lothian, Midlothian and West Lothian.
- Also in contrast to the geographic division, the dark blue Scottish cluster in the morphosyntax map includes both Dumfriesshire as well as Northumberland, a dialect site that is actually located in political England.
- In the morphosyntax partitioning, the red northern England area also comprises Shropshire and Leicestershire in what is often

referred to as the English Midlands, as well as Glamorganshire in southern Wales.

- Compared to the geographic map, the greenish southern English area is smaller in the morphosyntax partitioning: in linguistic terms, Shropshire, Leicestershire and Glamorganshire are – as we have seen – red dialects, while Middlesex and Warwickshire are yellow outliers. Durham, a county that is geographically located in northern England, is grouped with the southern English English dialects.

The morphosyntax dendrogram in Figure 6 demonstrates that by far the most fundamental split in the dataset occurs between English English dialects (red and green) and other dialects (including yellow outliers). The second most crucial split is between northern English English dialects (red) and southern English English dialects (green). The least important split is the one between central Scottish Lowland dialects (light blue) and other Scottish Lowland dialects (dark blue).

The name of the game in this section was classification. In this spirit, I have discussed ways to categorise dialects into discrete clusters. The case study suggests that despite some geographic incoherence and the presence of outliers, Great Britain can be divided into three major morphosyntactic dialect areas: the Scottish Lowlands *versus* the north of England *versus* the south of England.

7.4 Quantifying the explanatory power of language-external predictors

Dialectometry is intrinsically quantitative, yet my discussion has so far relied heavily on interpreting cartographic projections to geography. In this section, I introduce techniques used to correlate language-external parameters with linguistic distances, for the sake of precisely quantifying the extent to which dialect distances are predictable from language-external factors. To this end, the analyst typically starts out with an $N \times N$ linguistic distance matrix and creates parallel language-external distance matrices – one for each predictor to be tested. In the simplest case, each of these language-external distance matrices is then correlated with the linguistic distance matrix by calculating, for example, a Pearson product-moment correlation coefficient. The language-external predictor that scores the highest coefficient is the best predictor of linguistic distances.

To exemplify, I revisit our dataset on dialect variability in Great Britain. We begin by correlating our 34×34 morphosyntactic distance matrix with three language external distance matrices:

1. As-The-Crow-Flies Distances. Using a trigonometry formula on the FRED county coordinates, it is computationally trivial to calculate

pair-wise as-the-crow-flies distances (these actually underlie the left-hand projection in Figure 4 and the cluster map in Figure 5).[9] A proxy for the likelihood of social contact, as-the-crow-flies distance is the most common geographic distance measure in the literature (for example, Goebl, 2001; Gooskens and Heeringa, 2004; Nerbonne *et al*., 1996; and Shackleton, 2007).

2. Least-Cost Travel Times. Speakers do not have wings, so it is reasonable to assume that what really matters for dialect distances is how much time it would take a human traveller to get from point A to point B (see Gooskens, 2005; and Szmrecsanyi, forthcoming). To calculate this measure, we turned to Google Maps,[10] which has a route finder tool that allows the user to enter longitude/latitude pairings for two locations to obtain a least-cost travel route and, crucially, an estimate of the total travel time. We queried Google Maps for all $34 \times 33/2 = 561$ dialect pairings, thus obtaining pair-wise least-cost-travel time estimates.[11]
3. Linguistic Gravity Indices. Trudgill (1974: 233) suggested a gravity model to account for geographic diffusion, claiming 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'. Using Trudgill's formula, I calculated linguistic gravity values for each of the 561 dialect pairings in the database, feeding in least-cost travel time as geographic distance measure and early twentieth-century population figures[12] (in thousands) as a proxy for speaker community size.

Figure 7 provides three scatterplots that graph morphosyntactic distances against the language-external distance measures. In all three cases, there is a highly significant relationship, and the direction of the effect is the theoretically expected one throughout: increasing as-the-crow-flies distance and increasing least-cost-travel time predict increasing morphosyntactic distance; conversely, increasing linguistic gravity indices predict decreasing morphosyntactic distance. The R^2 values reported in Figure 7 suggest that

[9] The *R*u*G*/*L04* dialectometry software package provides a module (*ll2dst*) that can do this job automatically.

[10] See: http://maps.google.co.uk/

[11] These estimates assume travel by car (Google Maps' 'walking' option yields a matrix with a substantially lower correlation with linguistic distances). I fully acknowledge that matching linguistic data sourced from speakers born around the beginning of the twentieth century with travel estimates based on twenty-first century transportation infrastructure is convenient but clearly suboptimal; what is really needed are historic travel estimates, which, alas, are in short supply. Still, I submit that the procedure is not fatally flawed, since it can be argued that modern infrastructure in fact follows, to a large extent, historical travel routes, trade patterns and avenues of social contact.

[12] Specifically, I used 1901 population 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/

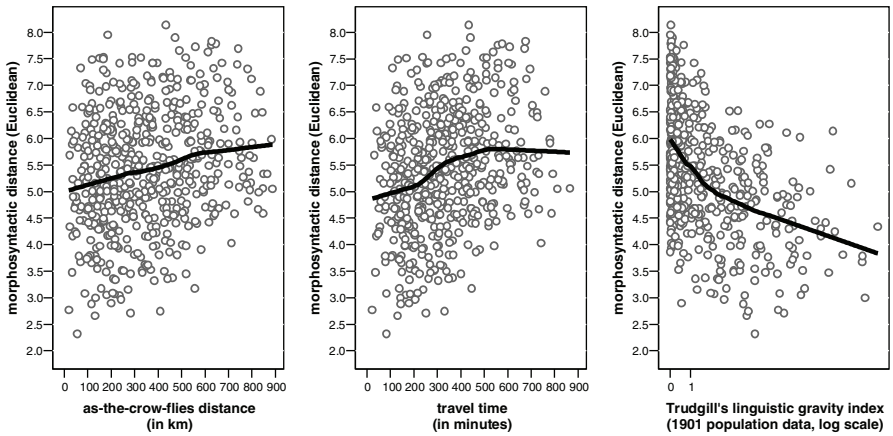

Figure 7: Correlating distance matrices: morphosyntactic distances *versus* (*i*) as-the-crow-flies distance (left) ($r = .21$, $p < .001$, $R^2 = 4.4$ percent), (*ii*) least-cost travel time (middle) ($r = .27$, $p < .001$, $R^2 = 7.4$ percent), and (*iii*) Trudgill's linguistic gravity index (right; log scale) ($r = -.49$, $p < .001$, $R^2 = 24.1$ percent; logarithmic estimate). Each dot represents one of $n = 561$ unique dialect pairings. Solid lines are LOESS curves estimating the overall nature of the relationship

as-the-crow-flies distance accounts for 4.4 percent of the morphosyntactic variance, least-cost travel time for 7.4 percent, and linguistic gravity for 24.1 percent. Hence, by factoring in speaker community size in addition to geographic distance, we can explain up to a quarter of the variance in morphosyntactic dialect distances. That Trudgill's linguistic gravity model turns out to be the most successful predictor of morphosyntactic distances in the dataset is interesting, given that some previous research has failed to detect a significant effect of linguistic gravity in, for example, Dutch dialects (see Heeringa *et al*., 2007; and Nerbonne and Heeringa, 2007). I conclude from this that, unlike in Dutch dialectology, the model works rather well for morphosyntactic variation in traditional British English dialects.

Having said that, however, I would wish to emphasise that in comparison to previous research the R^2 values reported here are rather low, anyway. For example, Shackleton (2007), in his study of phonetic variation in the Survey of English Dialects, reports R^2 values of up to 66 percent for the relationship between phonetic and geographic distances in England; Spruit *et al*. (2009), in an atlas-based study on aggregate syntactic distances in Dutch dialects, calculate an R^2 value of 45 percent for the relation between syntax and geography. So, given that the continuous-distance measures in Figure 7 somewhat fail us, could it be that the dialect partition I presented under Section 7.3 is a more potent predictor of morphosyntactic dialect distances? The problem is that binary dialect area (or dialect grouping) memberships do not elegantly lend themselves to a straightforward

continuous quantification with which one may furnish a full-blown $N \times N$ distance matrix that could be correlated with the original dialect distance matrix and depicted in a scatterplot. This is why we must resort to a technique known as Permutational Multivariate Analysis of Variance Using Distance Matrices (PERMANOVA; see Anderson, 2001)[13], which is analogous to MANOVA (Multivariate Analysis of Variance) but designed specifically to analyse distance matrices. The goal was to test how successful dialect-area membership is in predicting continuous dialect distances between measuring points. It turns out that the five-cluster partition depicted in Figure 6 explains about one third of the variance in morphosyntactic distances ($R^2 = 35.9$ percent, $p = .001$), which is a good deal more than we can explain by regressing as-the-crow-flies distance, travel time or linguistic gravity against morphosyntactic distances.

This section was an exercise in number crunching, and we have seen that the relationship between dialect distances and properties of geographic space is amenable to fairly precise quantification. In traditional British English dialects, as-the-crow-flies distance turns out to be a fairly poor predictor of aggregate morphosyntactic distances, explaining no more than about 4 percent of the overall variability. By factoring in parameters such as size of speaker community and travel distance, the analyst can boost the share of the linguistic variability that is accounted for to about 25 percent. At the same time, we have also seen that the cluster analytic dialect partition presented in the previous section explains more than a third of the morphosyntactic variance in the dataset. In other words, the dialect area scenario appears to be more appropriate for our dataset than the dialect continuum scenario.

8. Conclusion

This paper has presented methodologies which can be used to combine corpus-based variation studies with aggregative-dialectometrical analysis and visualisation methods. I have argued that this synthesis is desirable for two principal reasons. First, multidimensional objects, such as dialects, call for aggregate analysis techniques; second, vis-à-vis linguistic atlas material, corpora yield an arguably more trustworthy frequency signal. To exemplify the empirical potential of corpus-based dialectometry, we have drawn on a major dialect corpus to study aggregate relations between thirty-four traditional British English dialects, on the basis of joint variability in text frequencies of fifty-seven morphosyntactic features. The analysis has demonstrated that linguistic variability between British English dialects demonstrably provides a geographic signal, and that this signal has a number of interesting facets.

[13] To conduct the analysis, I utilised the statistical software package R: library *vegan*, function *adonis*. See: http://vegan.r-forge.r-project.org/

Needless to say, the line of analysis sketched in this paper is extendable in many ways. First and foremost, due to the case-study character of the investigation, we often stopped where real interpretation would start. More in-depth scholarship would seek to interpret the findings uncovered in this paper in a wider analytical and theoretical context, considering, among other things, the literature on dialect genesis, dialect formation, and historical dialect variability; theories on the status of, and constraints on, dialect grammar and grammatical variability between dialects from a typological perspective; and previous research on the issue of the contact-induced diffusibility *versus* universality of grammatical features.

Lastly, I should add that the methodology outlined in this paper is, of course, not limited to morphosyntactic phenomena. Phonology, lexis and even pragmatics are all in principle amenable to dialectometrical analysis using a corpus-based approach. There is even the intriguing possibility of aggregating not 'surfacy' feature frequencies but 'deep' feature conditionings (e.g., through probabilistic regression weights), which is something that simply cannot be done on the basis of decontextualised atlas or dictionary data. As for suitable databases, corpus-based dialectometry can be applied to any corpus in which we find geographic variability. This includes not only dialect corpora in the traditional sense (such as FRED, which is analysed here), but also corpora sampling geographically non-contiguous regional language varieties (such as the International Corpus of English; see Greenbaum, 1996) or corpora concerned with variation in written, not spoken, language (such as the letters-to-the-editor corpus presented in Grieve, 2009). In short, there are a great many research opportunities waiting to be tapped.

Acknowledgements

I am grateful to Hans Goebl, Wilbert Heeringa, Bernd Kortmann, John Nerbonne, Christoph Wolk and an anonymous referee for helpful comments and suggestions that have greatly improved this paper. The usual disclaimers apply.

Supplementary material

Larger versions of Figures 3, 4, 5 and 6 are available online as supplementary material to this article.

References

Aldenderfer, M.S. and R.K. Blashfield. 1984. Cluster Analysis. Newbury Park, London and New Delhi: Sage Publications.

Anderson, M.J. 2001. 'A new method for non-parametric multivariate analysis of variance', Austral Ecology 26, pp. 32–46.

Anderwald, L. and B. Szmrecsanyi. 2009. 'Corpus linguistics and dialectology' in A. Lüdeling and M. Kytö (eds) Corpus Linguistics: An International Handbook, pp. 1126–39. Berlin and New York: Mouton de Gruyter.

Arppe, A., G. Gilquin, D. Glynn, M. Hilpert and A. Zeschel. 2010. 'Cognitive Corpus Linguistics: five points of debate on current theory and methodology', Corpora 5 (2), pp. 1–27.

Auer, P., P. Baumann and C. Schwarz. Forthcoming. 'Vertical vs horizontal change in the traditional dialects of southwest Germany: a quantitative approach', Taal en Tongval.

Bland, J.M. and D.G. Altman. 1997. 'Statistics notes: Cronbach's alpha', British Medical Journal 314, p. 572.

Bloomfield, L. 1984 [1933]. Language. Chicago: University of Chicago Press.

Chafe, W. 1992. 'The importance of corpus linguistics to understanding the nature of language' in J. Svartvik (ed) Directions in Corpus Linguistics, pp. 79–97. Berlin and New York: Mouton de Gruyter.

Chambers, J.K. and P. Trudgill. 1998. Dialectology. Cambridge and New York: Cambridge University Press.

Cronbach, L.J. 1951. 'Coefficient alpha and the internal structure of tests', Psychometrika 16 (3), pp. 297–334.

Ellis, A.J. 1889. The Existing Phonology of English Dialects Compared with that of West Saxon Speech, Vol. V. London: Trübner and Co.

Embleton, S. 1993. 'Multidimensional scaling as a dialectometrical technique: outline of a research project' in R. Köhler and B. Rieger (eds) Contributions to Quantitative Linguistics, pp. 267–76. Dordrecht: Kluwer.

Giles, D. 2002. Advanced Research Methods in Psychology. Hove, New York: Routledge.

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

Goebl, H. 1984. Dialektometrische Studien: Anhand italoromanischer, rätroromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. Tübingen: Niemeyer.

Goebl, H. 2001. 'Arealtypologie und Dialektologie' in M. Haspelmath, E. König, W. Oesterreicher and W. Raible (eds) Language Typology and Language Universals/La typologie des langues et les universaux linguistiques/Sprachtypologie und sprachliche Universalien: An International Handbook/Manuel international/Ein internationales Handbuch, pp. 1471–91. Berlin and New York: Walter de Gruyter.

Goebl, H. 2005. 'La dialectométrie corrélative. Un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme', Revue de Linguistique Romane 69, pp. 321–67.

Goebl, H. 2006. 'Recent advances in Salzburg dialectometry', Literary and Linguistic Computing 21 (4), pp. 411–35.

Goebl, H. 2007. 'A bunch of dialectometric flowers: a brief introduction to dialectometry' in U. Smit, S. Dollinger, J. Hüttner, G. Kaltenböck and U. Lutzky (eds) Tracing English Through Time: Explorations in Language Variation, pp. 133–72. Wien: Braumüller.

Goebl, H. 2008. 'Le Laboratoire de dialectométrie de l'Université de Salzbourg', Zeitschrift für französische Sprache und Literatur 118 (1), pp. 35–55.

Goebl, H. and G. Schiltz. 1997. 'A dialectometrical compilation of CLAE 1 and CLAE 2: isoglosses and dialect integration' in W. Viereck and H. Ramisch (eds) Computer Developed Linguistic Atlas of England (CLAE), pp. 13–21. Tübingen: Max Niemeyer Verlag.

Gooskens, C. 2005. 'Traveling time as a predictor of linguistic distance', Dialectologia et Geolinguistica 13, pp. 38–62.

Gooskens, C. and W. Heeringa. 2004. 'Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data', Language Variation and Change 16 (3), pp. 189–207.

Greenbaum, S. 1996. Comparing English Worldwide: The International Corpus of English. Oxford and New York: Clarendon Press.

Gries, St. Th. and S. Wulff. 2005. 'Do foreign language learners also have constructions? Evidence from priming, sorting, and corpora', Annual Review of Cognitive Linguistics 3, pp. 182–200.

Grieve, J. 2009. A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English. Ph.D. Thesis, Northern Arizona University.

Haimerl, E. 2006. 'Database design and technical solutions for the management, calculation, and visualization of dialect mass data', Literary and Linguistic Computing 21 (4), pp. 437–44.

Haspelmath, M. 2009. 'Welche Fragen können wir mit herkömmlichen Daten beantworten?', Zeitschrift für Sprachwissenschaft 28 (1), pp. 157–62.

Heeringa, W. 2004. Measuring Dialect Pronunciation Differences Using Levenshtein Distance. Ph.D. Thesis. University of Groningen.

Heeringa, W., K. Johnson and C. Gooskens. 2009. 'Measuring Norwegian dialect distances using acoustic features', Speech Communication 51 (2), pp. 167–83.

Heeringa, W. and J. Nerbonne. 2001. 'Dialect areas and dialect continua', Language Variation and Change 13 (3), pp. 375–400.

Heeringa, W., J. Nerbonne, R. van Bezooijen and M. René Spruit. 2007. 'Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied', Nederlandse Taal-en Letterkunde 123 (1), pp. 70–82.

Hernández, N. 2006. User's Guide to FRED. Freiburg: University of Freiburg. URN: urn:nbn:de:bsz:25-opus-24895. Available online at: http: //www.freidok.uni-freiburg.de/volltexte/2489/

Holman, E.W., C. Schulze, D. Stauffer and S. Wichmann. 2007. 'On the relation between structural diversity and geographical distance among languages: observations and computer simulations', Linguistic Typology 11 (2), pp. 393–421.

Hoppenbrouwers, C. and G. Hoppenbrouwers. 2001. De indeling van de Nederlandse streektalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM. Assen: Koninklijke Van Gorcum.

Jain, A.K., M. Narasimha Murty and P.J. Flynn. 1999. 'Data clustering: a review', ACM Computing Surveys 31 (3), pp. 264–323.

Kortmann, B. and B. Szmrecsanyi. 2004. 'Global synopsis: morphological and syntactic variation in English' in B. Kortmann, E. Schneider, K. Burridge, R. Mesthrie and C. Upton (eds) A Handbook of Varieties of English, pp. 1142–202. Berlin and New York: Mouton de Gruyter.

Kruskal, J.B. and M. Wish. 1978. Multidimensional Scaling. Vol. 11. Newbury Park, London, New Delhi: Sage Publications.

Leech, G.N., B. Francis and X. Xu. 1994. 'The use of computer corpora in the textual demonstrability of gradience in linguistic categories' in C. Fuchs and B. Victorri (eds) Continuity in Linguistic Semantics, pp. 57–76. Amsterdam and Philadelphia: Benjamins.

Leinonen, T. 2008. 'Factor analysis of vowel pronunciation in Swedish dialects', International Journal of Humanities and Arts Computing 2 (1–2), pp. 189–204.

Meyers, L.S., G. Gamst and A.J. Guarino. 2006. Applied Multivariate Research: Design and Interpretation. Thousand Oaks: Sage Publications.

Nerbonne, J. 2006. 'Identifying linguistic structure in aggregate comparison', Literary and Linguistic Computing 21 (4), pp. 463–75.

Nerbonne, J. 2009. 'Data-driven dialectology', Language and Linguistics Compass 3 (1), pp. 175–98.

Nerbonne, J. and W. Heeringa. 1997. 'Measuring dialect distance phonetically' in J. Coleman (ed.) Workshop on Computational Phonology, Special Interest Group of the Association for Computational Linguistics, pp. 11–18. Madrid.

Nerbonne, J. and W. Heeringa. 2007. 'Geographic distributions of linguistic variation reflect dynamics of differentiation' in S. Featherston and W. Sternefeld (eds) Roots: Linguistics in Search of its Evidential Base, pp. 267–97. Berlin and New York: Mouton de Gruyter.

Nerbonne, J. and P. Kleiweg. 2007. 'Toward a dialectological yardstick', Journal of Quantitative Linguistics 14 (2), pp. 148–66.

Nerbonne, J., W. Heeringa and P. Kleiweg. 1999. 'Edit distance and dialect proximity' in D. Sankoff and J. Kruskal (eds) Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, pp. v–xv. Stanford: CSLI Press.

Nerbonne, J., W. Heeringa, E. van den Hout, P. van de Kooi, S. Otten and W. van de Vis. 1996. 'Phonetic distance between Dutch dialects' in G. Durieux, W. Daelemans and S. Gillis (eds) CLIN VI: Proceedings of the Sixth CLIN Meeting, pp. 185–202. Antwerp: Centre for Dutch Language and Speech (UIA).

Nerbonne, J., P. Kleiweg and F. Manni. 2008. 'Projecting dialect differences to geography: bootstrapping clustering vs. clustering with noise' in C. Preisach, L. Schmidt-Thieme, H. Burkhardt and R. Decker (eds) Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society, pp. 647–54. Berlin: Springer.

Nishisato, S. 2007. Multidimensional Nonlinear Descriptive Analysis. Boca Raton: Chapman and Hall/CRC.

Nunnally, J.C. 1978. Psychometric Theory. New York: McGraw-Hill.

Orton, H., S. Sanderson and J.D.A. Widdowson. 1978. The Linguistic Atlas of England. London and Atlantic Highlands, New Jersey: Croom Helm.

Penhallurick, R. 1993. 'Welsh English: a national language?', Dialectologia et Geolinguistica 1993 (1), pp. 28–46.

Penhallurick, R. 2004. 'Welsh English: morphology and syntax' in B. Kortmann, E. Schneider, K. Burridge, R. Mesthrie and C. Upton (eds) A Handbook of Varieties of English, pp. 102–13. Berlin and New York: Mouton de Gruyter.

Penke, M. and A. Rosenbach. 2007. 'What counts as evidence in linguistics? An introduction', Studies in Language 28 (3), pp. 480–526.

Séguy, J. 1971. 'La relation entre la distance spatiale et la distance lexicale', Revue de Linguistique Romane 35, pp. 335–57.

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

Speelman, D. and D. Geeraerts. 2008. 'The role of concept characteristics in lexical dialectometry', International Journal of Humanities and Arts Computing 2 (1–2), pp. 221–42.

Spruit, M.R. 2005. 'Classifying Dutch dialects using a syntactic measure: the perceptual Daan and Blok dialect map revisited', Linguistics in the Netherlands 22 (1), pp. 179–90.

Spruit, M.R. 2006. 'Measuring syntactic variation in Dutch dialects', Literary and Linguistic Computing 21 (4), pp. 493–506.

Spruit, M.R., W. Heeringa and J. Nerbonne. 2009. 'Associations among linguistic levels', Lingua 119 (11), pp. 1624–42.

Szmrecsanyi, B. 2008. 'Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects', International Journal of Humanities and Arts Computing 2 (1–2), pp. 279–96.

Szmrecsanyi, B. 2010. The Morphosyntax of BrE Dialects in a Corpus-based Dialectometrical Perspective: Feature Extraction, Coding Protocols, Projections to Geography, Summary Statistics. Freiburg: University of Freiburg. URN: urn:nbn:de:bsz:25-opus-73209. Available online at: http://www.freidok.uni-freiburg.de/Volltexte/7320/

Szmrecsanyi, B. Forthcoming. 'Geography is overrated' in S. Hansen, C. Schwarz, P. Stoeckle and T. Streck (eds) Dialectological and Folk Dialectological Concepts of Space. Berlin: Walter de Gruyter.

Szmrecsanyi, B. and N. Hernández. 2007. Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S"). Freiburg: University of Freiburg. URN: urn:nbn:de:bsz: 25-opus-28598. Available online at: http://www.freidok.uni-freiburg.de/volltexte/2859/

Trudgill, P. 1974. 'Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography', Language in Society 3 (2), pp. 215–46.

Trudgill, P. 1999. The Dialects of England. Cambridge, Massachusetts and Oxford: Blackwell.

Viereck, W. 1985. 'Linguistic atlases and dialectometry: the survey of English dialects' in J.M. Kirk, S. Sanderson and J.D.A. Widdowson (eds) Studies in Linguistic Geography: The Dialects of English in Britain and Ireland, pp. 94–112. London: Croom Helm.

Viereck, W., H. Ramisch, H. Händler, P. Hoffmann and W. Putschke. 1991. The Computer Developed Linguistic Atlas of England. Tübingen: Niemeyer.

Voronoi, G. 1907. 'Nouvelles applications des paramètres continus à la théorie des formes quadratiques', Journal für die Reine und Angewandte Mathematik 133, pp. 97–178.

Wälchli, B. 2009. 'Data reduction typology and the bimodal distribution bias', Linguistic Typology 13 (1), pp. 77–94.

Ward, J.H. Jr. 1963. 'Hierarchical grouping to optimize an objective function', Journal of the American Statistical Association 58, pp. 236–44.

Appendix A: The feature catalogue

A. Pronouns and determiners

(1) non-standard reflexives (e.g., *they didn't go theirself*)
(2) standard reflexives (e.g., *they didn't go themselves*)
(3) archaic *thee*/*thou*/*thy* (e.g., *I tell thee a bit more*)
(4) archaic *ye* (e.g., *ye'd dancing every week*)
(5) *us* (e.g., *us couldn't get back*, *there was no train*)
(6) *them* (e.g., *I wonder if they'd do any of them things today*)

B. The noun phrase

(7) synthetic adjective comparison (e.g., *he was always keener on farming*)
(8) the *of* -genitive (e.g., *the presence of my father*)
(9) the *s*-genitive (e.g., *my father's presence*)
(10) preposition stranding (e.g., *the very house which it was in*)
(11) cardinal number + *years* (e.g., *I was there about three years*)
(12) cardinal number + *year*-Ø (e.g., *she were three year old*)

C. Primary verbs

(13) the primary verb TO DO (e.g., *why did you not wait?*)
(14) the primary verb TO BE (e.g., *I was took straight into this pitting job*)
(15) the primary verb TO HAVE (e.g., *we thought somebody had brought them*)
(16) marking of possession, HAVE GOT (e.g., *I have got the photographs*)

D. Tense and aspect

(17) the future marker BE GOING TO (e.g., *I'm going to let you into a secret*)
(18) the future markers WILL/SHALL (e.g., *I will let you into a secret*)
(19) WOULD as marker of habitual past (e.g., *he would go around killing pigs*)
(20) *used to* as marker of habitual past (e.g., *he used to go around killing pigs*)
(21) progressive verb forms (e.g., *the rest are going to Portree School*)
(22) the present perfect with auxiliary BE (e.g., *I'm come down to pay the rent*)
(23) the present perfect with auxiliary HAVE (e.g., they'*ve killed* the skipper)

Appendix A (*continued*): The feature catalogue

E. Modality

(24) marking of epistemic and deontic modality: MUST (e.g., *I must pick up the book*)
(25) marking of epistemic and deontic modality: HAVE TO (e.g., *I have to pick up the book*)
(26) marking of epistemic and deontic modality: GOT TO (e.g., *I gotta pick up the book*)

F. Verb morphology

(27) a-prefixing on *–ing*-forms (e.g., *he was a-waiting*)
(28) non-standard weak past tense and past participle forms (e.g., *they knowed all about these things*)
(29) non-standard past tense *done* (e.g., *you came home and done the home fishing*)
(30) non-standard past tense *come* (e.g., *he come down the road one day*)

G. Negation

(31) the negative suffix *–nae* (e.g., *I cannae do it*)
(32) the negator *ain't* (e.g., *people ain't got no money*)
(33) multiple negation (e.g., *don't you make no damn mistake*)
(34) negative contraction (e.g., *they won't do anything*)
(35) auxiliary contraction (e.g., *they'll not do anything*)
(36) *never* as past tense negator (e.g., *and they never moved no more*)
(37) WASN'T (e.g., *they wasn't hungry*)
(38) WEREN'T (e.g., *they weren't hungry*)

H. Agreement

(39) non-standard verbal *–s* (e.g., *so I says*, *What have you to do?*)
(40) *don't* with third person singular subjects (e.g., *if this man don't come up to it*)
(41) standard *doesn't* with third person singular subjects (e.g., *if this man doesn't come up to it*)
(42) existential/presentational *there is*/*was* with plural subjects (e.g., *there was children involved*)
(43) absence of auxiliary BE in progressive constructions (e.g., *I said*, *How Ø you doing?*)
(44) non-standard WAS (e.g., *three of them was killed*)
(45) non-standard WERE (e.g., *he were a young lad*)

Appendix A (*continued*): The feature catalogue

I. Relativisation

(46) *wh*-relativisation (e.g., *the man who read the book*)
(47) the relative particle *what* (e.g., *the man what read the book*)
(48) the relative particle *that* (e.g., *the man that read the book*)

J. Complementation

(49) *as what* or *than what* in comparative clauses (e.g., *we done no more than what other kids used to do*)
(50) unsplit *for to* (e.g., *it was ready for to go away with the order*)
(51) infinitival complementation after BEGIN, START, CONTINUE, HATE and LOVE (e.g., *I began to take an interest*)
(52) gerundial complementation after BEGIN, START, CONTINUE, HATE and LOVE (e.g., *I began taking an interest*)
(53) zero complementation after THINK, SAY, and KNOW (e.g., *they just thought Ø it isn't for girls*)
(54) that complementation after THINK, SAY and KNOW (e.g., *they just thought that it isn't for girls*)

K. Word order and discourse phenomena

(55) lack of inversion and/or of auxiliaries in *wh*-questions and in main clause *yes*/*no*-questions (e.g., *where you put the shovel*?)
(56) the prepositional dative after the verb GIVE (e.g., *she gave* [*a job*] *to* [*my brother*])
(57) double object structures after the verb GIVE (e.g., *she gave* [*my brother*] [*a job*])