Corpus-based dialectometry: aggregate morphosyntactic variability in British English dialects*

Benedikt Szmrecsanyi
Freiburg Institute for Advanced Studies
bszm@frias.uni-freiburg.de

Abstract

The research reported in this paper departs from most previous work in dialectometry in several ways. Empirically, it draws on frequency vectors derived from naturalistic corpus data and not on discrete atlas classifications. Linguistically, it is concerned with morphosyntactic (as opposed to lexical or pronunciational) variability. Methodologically, it marries the careful analysis of dialect phenomena in authentic, naturalistic texts to aggregational-dialectometrical techniques. Two research questions guide the investigation: First, on methodological grounds, is corpus-based dialectometry viable at all? Second, to what extent is morphosyntactic variation in non-standard British dialects patterned geographically? By way of validation, findings will be matched against previous work on the dialect geography of Great Britain.

1 Introduction

The overarching aim in this study is to provide a methodological sketch of how to blend philologically responsible corpus-based research with aggregational-dialectometrical analysis techniques. The bulk of previous research in dialectometry has focussed on phonology and lexis (however, for work on Dutch dialect syntax see Spruit 2005, 2006, 2008, Spruit et al. t.a.). Moreover, orthodox dialectometry draws on linguistic atlas classifications as its primary data source. The present study departs from these traditions in several ways. It endeavours, first, to measure aggregate *morphosyntactic* distances and similarities between traditional dialects in the British Isles. Second, the present study does not rely on atlas data but on frequency information deriving from a careful analysis of language use in authentic, naturalistic texts. This is another way of saying that the aggregate analysis in this paper is *frequency-based*, an approach that contrasts with *atlas-based* dialectometry, which essentially relies on categorical input data. Succinctly put, the difference is that atlas-based approaches typically aggregate observations such as *of two variants X and Y, variant X is the dominant one in dialect Z*, while frequency-based approaches are empirically based on corpus findings along the lines of, say, *in dialect Z, variant X is 3.5 times more frequent in actual speech than variant Y*.

The corpus resource drawn on is FRED, the *Freiburg English Dialect Corpus*, a naturalistic speech corpus sampling interview material from 162 different locations in 38 different counties all over the British Isles, excluding Ireland. The corpus was analyzed to obtain text frequencies of 62 morphosyntactic features, yielding a structured database that provides a 62-dimensional frequency vector per locality. The Euclidean distance measure was subsequently applied to compute aggregate morphosyntactic distances, which then served as the input to dialectometrical analysis.

Two research questions guide the present study's inquiry: first, on the methodological plane we are interested in whether and how corpus-based (that is, frequency-based) dialectometry is viable. Substantially, we will seek to uncover if and to what extent morphosyntactic variation in non-standard British dialects is patterned along geographic lines. By way of validation, findings will be matched against previous work (dialectological, dialectometrical, and perceptual) on the dialect geography of Great Britain.

2 Previous work on aggregate dialect differences in Great Britain

Let us first turn to the literature in order to eclectically review extant scholarship on dialect differences in Great Britain. **?**:20–35 is one of the best-known dialectological accounts of accent differences in traditional British dialects. **?** studies eight salient accent features to establish a composite map dividing England into 13 traditional dialect areas. These can be grouped into six macro areas: (1) *Scots*, (2) *northern dialects* (Northumberland and the Lower North), (3) *western central (Midlands) dialects* (Lancashire, Staffordshire), (4) *eastern central (Midlands) dialects* (South Yorkshire, Lincolnshire, Leicestershire), (5) *southwestern dialects* (western Southwest, northern Southwest, eastern Southwest), and (6) *southeastern dialects* (central East and eastern Countries).

In the realm of perceptual dialectology, Inoue (1996) conducted an experiment to study the subjective dialect division in Great Britain. 77 students at several universities in Great Britain were asked, among other things, to draw lines on a blank map 'according to the accents or dialects they perceived' (Inoue 1996:146), based on their experience. The result of this exercise can be summarised as follows: dialects of English in Wales and Scotland are perceived as being very different from English English dialects. Within England, the North is differentiated from the Midlands, and the Midlands are differentiated from the South (Inoue 1996:map 3). This division is quite compatible with **?**'s (**?**) classification, except that in Inoue's (1996) experiment, Lancashire is part of the North, not of the western Midlands, and the northern Southwest (essentially, Shropshire and Herfordshire) patterns with Midland dialects, not southwestern dialects.

As for atlas-based dialectometry, Goebl (2007) draws on the *Computer Developed Linguistic Atlas of England* (which is based on the *Survey of English Dialects*) to study aggregate linguistic relationships between 314 sites all over England. The aggregate analysis is based on 597 lexical and morphosyntactic features. Among many other things, Goebl (2007) utilises cluster analysis to partition England into discrete dialect areas (Goebl 2007:maps 17–18). It turns

out that there is 'a basic opposition between the North [...] and the South of England' (Goebl 2007:145). The dividing line runs south of Lancashire and South Yorkshire, and thus cuts right across what **?** and Inoue (1996) classify as the Midlands dialect area. In southern English dialects, Goebl (2007) finds a major split between southwestern and other southern dialects.

3 Methods and data

The present study is an exercise in corpus-based dialectometry. Corpus linguistics is a methodology that draws on principled collections of naturalistic texts to explore authentic language usage. A hallmark of the methodology is the 'extensive use of computers for analysis, using both automatic and interactive techniques' and the reliance 'on both quantitative and qualitative analytical techniques' (Biber et al. 1998:4). This section will discuss the corpus as well as the feature frequency portfolio that will serve as the basis for the subsequent aggregate analysis.

3.1 Data source: the *Freiburg English Dialect Corpus* (FRED)

This study will tap the *Freiburg English Dialect Corpus* (henceforth: FRED) (see Hernández 2006; Szmrecsanyi and Hernández 2007 for manuals) as its primary data source. FRED contains 372 individual texts and spans approximately 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. Most of these samples were recorded between 1970 and 1990; in most cases, a field-worker interviewed 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 (so-called 'non-mobile old rural males'). 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 corpus is annotated with longitude/latitude information for each of the locations sampled. From this annotation, county coordinates can be calculated by computing the arithmetic mean of all the location coordinates associated with a particular county. At present, FRED is neither part-of-speech annotated nor syntactically parsed.

3.2 Feature selection and extraction

Corpus-based dialectometry is essentially frequency-based dialectometry; thus the approach outlined here bears a certain similarity to the method in Hoppenbrouwers and Hoppenbrouwers (2001) (discussed in Heeringa 2004:16–20). Following a broadly variationist approach in the spirit of, for example, Labov (1966), a catalogue spanning 35 morphosyntactic variables with typically (but not always) two variants each was defined. This catalogue of 35 variables yields a list of $p = 62$ morphosyntactic target variants (henceforth: *features*); the Appendix provides a comprehensive list. In an attempt to aggregate as many variables as possible, the features included in the catalogue are the usual suspects in the dialectological, variationist, and corpus-linguistic literature, regardless of

whether a geographic distribution has previously been reported for a particular feature or not. To qualify for inclusion, however, a candidate feature had to fulfill the following criteria:

1. For statistical reasons, the feature had to be relatively frequent, specifically: ≥ 1 occurrence per 10,000 words of running text (this criterion rules out interesting but infrequent phenomena such as resumptive relative pronouns or double modals).

2. For practical purposes, the feature had to be extractable subject to a reasonable input of labour resources by a human coder (ruling out, for example, hard-to-retrieve null phenomena such as zero relativisation, or phenomena where semantics enters heavily into consideration, such as gendered pronouns).

Next, the material in FRED was coded for the features in the catalogue. 26 features for which automatic recall was feasible were extracted automatically using Perl (*Practical Extraction and Report Language*) scripts. 36 features were coded manually after pre-screening the data using Perl scripts, a step which considerably narrowed down the number of phenomena which had to be inspected manually. Even so, the frequency database utilised in the present study is based on 75,124 manual (that is, qualitative) coding decisions. Szmrecsanyi (forthcoming) provides a detailed description of the procedure along with the detailed coding schemes that regimented the coding process.

Once coding was complete, another line of Perl scripts was used to extract vectors of $p_{total} = 62$ feature frequencies per locality. The feature frequencies were subsequently normalised to *frequency per ten thousand words* (because textual coverage in FRED varies across localities) and *log*-transformed* to deemphasise large frequency differentials and to alleviate the effect of frequency outliers. The resulting 38×62 table (on the county level – that is, 38 counties characterised by 62 feature frequencies each for the full dataset) yields a Cronbach's α value of .86, indicating satisfactory reliability.

Finally, the 38×62 table was converted into a 38×38 distance matrix using Euclidean distance – the square root of the sum of all squared frequency differentials – as an interval measure. This distance matrix was subsequently analyzed dialectometrically.*

4 Results

We now move on to a discussion of empirical findings. Unless stated otherwise, the level of areal granularity is the county level ($N = 38$).

4.1 On the explanatory power of geography

Let us first consider the role that geographic distance plays in aggregate morphosyntactic variability. First, how much of this variability can be explained by geography? Second, looking at the morphosyntactic dialect landscape in the British Isles, to what extent are we dealing with a continuum such that transitions are gradual and not abrupt?

As for the first question, a Perl script was run on the Euclidean distance matrix based on all $p_{total} = 62$ features and on FRED's geographic longitude/latitude annotation to generate a table specifying pairwise morphosyntactic and geographic distances. This yielded an exhaustive list of all $N \times \frac{N-1}{2} = 703$ possible county pairings, each pairing being annotated for morphosyntactic and geographic distance. On the basis of this list, the scatterplot in Figure 1 illustrates the correlation between morphosyntactic and geographic distance in the database at hand.

[Figures 1 and 2 here]

Figure 1 highlights two facts. First, while the correlation between morphosyntactic and geographic distance is highly significant ($p = .00$), it is relatively weak (Pearson correlation coefficient: $r = .22$). In other words, geography explains overall only 4.7 per cent of the morphosyntactic variance ($R^2 = .047$). To put this value into perspective, Spruit et al. (to appear:Table 7) – in a study on aggregate linguistic distances in Dutch dialects – report R^2 values of .47 for the correlation between geography and pronunciation, .33 for lexis, and .45 for syntax. Second, the best curve estimation for the relationship between morphosyntactic and geographic distance in British English dialects is actually linear.* Given Séguy (1971) and much of the atlas-based dialectometry literature that has followed Séguy's seminal study, one would actually expect a sublinear or logarithmic relationship. Having said that, we note that Spruit (2008:54-55), in his study of Dutch dialects, finds that the correlation between syntactic and geographic distance is also more linear than logarithmic. Hence, it may simply be the case that (morpho)syntactic variability has a different relationship to geographic distance than lexical or pronunciational variability.

Against this backdrop, it is interesting to note that not all of the 62 features entered into aggregate analysis correlate significantly with geography. In fact, only 23 features do (these are marked with an asterisk in the Appendix).* When the aggregate analysis is based on only those $p_{geo} = 23$ features, we obtain the scatterplot in Figure 2. The correlation coefficient between morphosyntactic and geographic distance is now approximately twice as high as in Figure 1 ($r = .41$), which means that for this particular feature subset geography explains about 16.6 per cent of the morphosyntactic variance ($R^2 = .166$).* While these numbers begin to approximate the explanatory potency of geography in atlas-based dialectometry, it still seems that we should base the aggregate analysis on all available data. This is why the subsequent analysis in this paper will be based on the entire feature portfolio ($p_{total} = 62$), despite the weaker geographic signal it provides. Still, we observe that feature selection does matter a great deal, and one is left to wonder to what extent compilers of linguistic atlases – the primary data source for those studies that report high coefficients for geography – really draw on all available features, or rather on those features that seem geographically interesting.

[Figure 3 here]

Comparatively weak as the overall correlation between morphosyntactic and geographic distance may be, are we nonetheless dealing with a morphosyntactic dialect continuum? To answer this question, we will now visualise aggregate morphosyntactic variability using cartographic techniques, all relying on Voronoi tesselation (see Goebl 1984) to project linguistic results to geography. Regular multidimensional scaling (henceforth: MDS) (see Kruskal and Wish 1978) was utilised to scale down the original 62-dimensional Euclidean distance matrix to three dimensions; the distances in the three-dimensional MDS solution correlate with the distances in the original distance matrix to a satisfactory degree ($r = .82$). Subsequently, the three MDS dimensions were mapped to the red–green–blue colour components, giving each of the county polygons in Figure 3 a distinct colour.* In continuum maps such as Figure 3, smooth (as opposed to abrupt) colour transitions implicate the presence of a dialect continuum. As can be seen, the morphosyntactic dialect landscape in the British Isles is overall not exceedingly continuum-like.* While colour transitions in the south of England are fairly smooth (meaning that this is a fairly homogeneous dialect area), the picture is more noisy in the North of England and, especially, in Scotland. To aid interpretation of Figure 3, each of the 62 normalised *log*-transformed feature frequencies was correlated against each of the three MDS dimensions to determine which of the features correlate most strongly with the red–green–blue colour scheme in Figure 3 (see Wieling et al. 2007 for a similar procedure). It turns out that more reddish colours correlate best with increased frequencies of multiple negation (feature [34]) ($r = .79$), greenish colours correlate most strongly with higher frequencies of non-standard weak past tense and past participle forms (feature [23]) ($r = .63$), and bluish colours correlate best with increased frequencies of *wh*-relativisation (feature [49]) ($r = .57$).

By way of an interim summary, the research discussed in this section has two principal findings. Firstly, the explanatory potency of geography is comparatively weak in the data at hand and accounts for only between 4.7 to 16.6 per cent of the observable morphosyntactic variance (depending on whether all available features or only those with a significant geographic distribution are studied). Secondly, the morphosyntactic dialect landscape in Great Britain does not have a very continuum-like structure overall, although transitions appear to be more gradual in England than in Scotland.

4.2 Classification and validation

The task before us now is to examine higher-order patterns and groupings among British English dialects. Is it possible to identify dialect areas on morphosyntactic grounds (and on the empirical basis of frequency data)? If so, do these dialect areas conform to those previously identified in the literature (see section 2)?

To answer these questions, hierarchical agglomerative cluster analysis (see Aldenderfer and Blashfield 1984), a data classification technique used to partition observations into discrete groups, was applied to the dataset. Simple clustering can be unstable, hence a procedure known as 'clustering with noise' (Nerbonne et al. 2008) was conducted: the original Euclidean distance matrix was clustered repeatedly, adding some random amount of noise in each run. This exercise yielded a cophenetic distance matrix which details consensus (and thus more stable) cophenetic distances between localities, and which is amenable

to various cartographic visualisation techniques. This study uses the clustering parameters described in Nerbonne et al. (2008), setting a noise ceiling of $c = \sigma/2$ and performing 100 clustering runs. There are many different clustering algorithms; in addition to using the – quite customary – *Weighted Pair Group Method using Arithmetic Averages* (WPGMA), we also apply *Ward's Minimum Variance Method* (WARD), as the two algorithms yield interestingly different clustering outcomes.*

[Figures 4, 5, 6, and 7 here]

The resulting higher-order structures can be visualised, for example, via so-called *composite cluster maps* (see Nerbonne et al. 2008 for a discussion). These highlight the fuzzy nature of dialect boundaries such that darker borders between localities represent more robust linguistic oppositions (which, thanks to the clustering-with-noise technique utilized, can be considered statistically significant). Figure 4 presents a composite cluster map that visualises the outcome of WPGMA noisy clustering, which is contrasted with the corresponding WARD outcome in Figure 5. An alternative visualisation, which highlights rough group memberships and fuzzy transition areas, can be attained by applying MDS to the cophenetic distance matrix (see, for instance, Alewijnse et al. 2007:section 5.3) and subsequently assigning component colours to each of the three resulting MDS dimensions. Such maps – where similar colourings indicate likely membership in the same dialect area – are displayed in Figure 6 (WPGMA) and Figure 7 (WARD). Note, in this context, that the distances in the three-dimensional MDS solution correlate very highly with the distances in the cophenetic distance matrix ($r = .96$ and $r = 1.00$, respectively).

Figures 4 through 7 can be interpreted as follows. Both the WPGMA and WARD algorithms characterise Scotland as heterogeneous and geographically fairly incoherent (more so according to WPGMA than according to WARD). Both algorithms moreover tend to differentiate between English English dialects and non-English English dialects (Scottish English dialects and northern Welsh dialects, in particular Denbighshire [DEN]). This is consonant with the sharp perceptual split between English English dialects and Welsh/Scottish dialects reported in Inoue (1996). As for divisions among English English dialects, however, the two clustering algorithms generate fairly different classifications:

- WPGMA classifies England as a rather homogeneous dialect area vis-à-vis Scotland and Wales. The only outlier in England is the county Warwickshire (WAR; the brownish polygon in Figure 6), which is more similar to Denbighsire (DEN; Welsh English) and some Scottish dialects than to the other English counties.

- WARD broadly distinguishes between southern English dialects (reddish/pinkish colours in Figure 7) and northern English dialects (brownish/darkish colours). Northumberland (NBL, dark green), Durham (DUR, blue), and Warwickshire (WAR; light blue), albeit English counties, pattern with Scottish dialects. Middlesex (MDS) is grouped with the northern dialects, although the county is located in the geographic Southeast (this fact is responsible for the salient southeastern 'box' in Figure 5). In sum, the WARD

algorithm finds a rather robust North–South split in England, which is compatible with all three accounts surveyed in Section 2 (**?**Inoue 1996; Goebl 2007). Figures 5 and 7 can also be seen to reveal a split among northern dialects into Midland dialects (darkish/brownish colours, in particular Leicestershire [LEI], Shropshire [SAL], Lancashire [LAN], Westmorland [WES], and Yorkshire [YKS]) versus northern dialects (Durham [DUR] and Northumberland [NBL]). This opposition would be in accordance with Inoue (1996) as well as **?**.

In summary, we have seen in this section that it seems to be possible – despite a good deal of apparent geographical incoherence – to identify rough dialect areas on morphosyntactic grounds, and that these are not incompatible with previous accounts of dialect differences in Great Britain. For one thing, most English English dialects are rather robustly differentiated from non-English English dialects. Second, the WARD algorithm in particular finds a North–South split among English English dialects that appears meaningful given extant scholarship. At the same time, we note that both algorithms fail to identify meaningful and coherent patterns among Scottish dialects. Also, neither algorithm detects a split between the Southwest of England and other southern dialects, as posited by **?** and Goebl (2007).

5 Conclusions

This study has demonstrated that frequency vectors derived from naturalistic corpus data – as opposed to, for instance, categorical linguistic atlas classifications – can serve as the empirical basis for aggregate analysis. Focussing on morphosyntactic variability in British English dialects, we have seen that the dataset yields a significant geographic signal which is, however, comparatively weak in comparison to previous atlas-based dialectometrical findings. The analysis has also suggested that overall variability in British English dialects does not seem to have an exceedingly continuum-like structure, and that there is quite a bit of geographical incoherence. Future study will want to investigate whether the comparatively weak explanatory potency of geography is real, or whether it is an artefact of the specific methodology or data type used. Having said that, the results do reveal that British English dialects can be partitioned into rough dialect areas on morphosyntactic grounds. Although the match with the literature is not perfect – as a matter of fact, we should not expect it to be perfect, given that some of the studies cited 'are based on entirely different things and on not very much at all', as one reviewer of this paper noted – the classification suggested here is not incompatible with previous work on dialect divisions in Great Britain. This enhances confidence in the method utilized here. A more detailed discussion of the outlier status of counties such as Warwickshire and Middlesex (including the identification of the features that are responsible for this outlier status), and of the extent to which the methodology presented here uncovers hitherto unknown generalisations is reserved for another occasion.

More generally speaking, though, the present study highlights the fact that a careful and philologically responsible identification and analysis of features occurring in naturalistic, authentic texts (as customary in, for example, variationist sociolinguistics and corpus-based dialectology) advertises itself for aggregation and computational analysis. The point is that the qualitative-philological

jeweller's eye perspective and the quantitative-aggregational bird's eye perspective are not mutually exclusive, but can be fruitfully combined to explore large-scale patterns and generalisations. It should be noted in this connection that the line of aggregate analysis sketched out in this paper could easily be extended to other humanities disciplines that rely on naturalistic texts as their primary data source (for instance, literary studies, historical studies, theology, and so on).

The methodology outlined in the present study can and should be refined in many ways. For one thing, work is under way to utilise Standard English text corpora to determine aggregate morphosyntactic distances between British English dialects, on the one hand, and standard English dialects (British and American) on the other hand. Second, the feature-based frequency information on which the present study rests will be supplemented in the near future by part-of-speech frequency information, on the basis of a coding scheme that distinguishes between 73 different part-of-speech categories. Third, given that geography does not seem to play an exceedingly important role in the dataset analyzed here, it will be instructive to draw on network diagrams (in the spirit of, for example, McMahon et al. 2007) as an additional visualisation and interpretation technique.

Notes

*I am grateful to John Nerbonne, Wilbert Heeringa, and Bart Alewijnse for having me over in Groningen in spring 2007 to explain dialectometry to me. I also wish to thank Peter Kleiweg for creating and maintaining the *RuG/L04* package. The audience at the Workshop on 'Measuring linguistic relations between closely related varieties' at the MethodsXIII conference in Leeds (August 2008) provided very helpful and valuable feedback on an earlier version of this paper, as did four anonymous reviewers. The usual disclaimers apply.

*Zero frequencies were rendered as .0001, which yields a *log* frequency of -4.

*The analysis was conducted using some custom-made Perl scripts, standard statistical software (SPSS), and Peter Kleiweg's *RuG/L04* package (available online at `http://www.let.rug.nl/~kleiweg/L04/`) as well as the L04 web interface maintained by Bart Alewijnse (`http://l04.knobs-dials.com/`).

*$R^2_{linear} = .0469$, $R^2_{logarithmic} = .0439$

*In order to test individual features for significant geographic distributions, dialect distances were also calculated on the basis of individual features (using one-dimensional Euclidean distance as interval measure) and correlated with geographical distance. If the ensuing correlation coefficient was significant, a given feature was classified as having a significant geographic distribution.

*Still, the relationship is more linear ($R^2_{linear} = .0166$) than logarithmic ($R^2_{logarithmic} = .134$).

*To do justice to FRED's areal coverage – which is unparalleled in the corpus-linguistic realm, but certainly not perfect – the polygons in Figure 3 have a maximum radius of ca. 40 km. This yields a 'patchy' but arguably more realistic geographic projection.

*Having said that, it should be made explicit that the present study is based on an aggregate analysis of features that are known to display variation (though not necessarily geographic variation). As one reviewer noted, the inclusion of more invariable features – say, basic word order or the like – would yield smoother dialect transitions. This is of course true, yet we note that linguistic atlases, and thus atlas-based dialectometry, also of course have a bias towards variable features.

*Notice that given the present study's dataset, the *Unweighted Pair Group Method using Arithmetic Averages* (UPGMA), another popular algorithm used in, for instance, Nerbonne et al. (2008), yields almost exactly the same classification as WPGMA.

Appendix: the feature catalogue

Features whose distribution correlates significantly with geography are marked by an asterisk (*).

A. The pronominal system

[1]*	vs.	[2]	non-standard vs. standard reflexives
[3]	vs.	[4]	archaic *thee*, *thou*, *thy* vs. standard *you, yours, you*

B. The noun phrase

[5]*	vs.	[6]	synthetic vs. analytic adjective comparison
[7]	vs.	[8]	the *of*-genitive vs. the *s*-genitive
[9]	vs.	[10]*	preposition stranding vs. preposition/particle frequencies

C. Primary verbs

[11]	vs.	[12]*	the primary verb TO DO vs. the primary verbs TO BE/HAVE NOTE: this includes both main verb and auxiliary verb usages

D. Tense, mood, and aspect

[13]	vs.	[14]	the future marker BE GOING TO vs. WILL/SHALL
[15]	vs.	[16]*	*would* vs. *used to* as markers of habitual past
[17]*	vs.	[18]	progressive vs. unmarked verb forms
[19]*	vs.	[20]	the present perfect with auxiliary BE vs. the present perfect with auxiliary HAVE

E. Verb morphology

[21]	vs.	[22]	*a*-prefixing on *-ing*-forms vs. bare *-ing*-forms
[23]	vs.	[24]	non-standard weak past tense and past participle forms vs. standard strong forms
[25]*	vs.	[26]	non-standard 'Bybee' verbs vs. corresponding standard forms NOTE: 'Bybee' verbs (see Anderwald 2009) have a three-way paradigm – e.g. *begin/began/begun* – in Standard English but can be reduced to a two-way paradigm – e.g. *begin/begun/begun* – in dialect speech
[27]			non-standard verbal *-s*
[28]*	vs.	[29]	non-standard past tense *done* vs. standard *did*
[30]	vs.	[31]	non-standard past tense *come* vs. standard *came*

F. Negation

[32]* vs. [33]	invariant *ain't* vs. *not/*n't/*nae*-negation
[34]* vs. [35]	multiple negation vs. simple negation
[36]* vs. [37]	negative contraction vs. auxiliary contraction
[38]* vs. [39]*	*don't* with 3rd person singular subjects vs. standard agreement
[40] vs. [41]	*never* as a preverbal past tense negator vs. standard negation

G. Agreement

[42]	existential/presentational *there is* vs. *was* with plural subjects
[43]* vs. [44]	deletion of auxiliary BE in progressive constructions vs. auxiliary BE present
[45]* vs. [46]*	non-standard WAS vs. standard WAS
[47] vs. [48]*	non-standard WERE vs. standard WERE

H. Relativisation

[49]	*wh*-relativisation
[50]*	relative particle *what*
[51]	relative particle *that*
[52]	relative particle *as*

I. Complementation

[53]*	*as what* or *than what* in comparative clauses
[54] vs. [55]*	unsplit *for to* vs. *to*-infinitives
[56] vs. [57]	infinitival vs. gerundial complementation after TO BEGIN, TO START, TO CONTINUE, TO HATE, TO LOVE
[58] vs. [59]	zero vs. *that* complementation after TO THINK, TO SAY, and TO KNOW

J. Word order phenomena

[60]	lack of inversion and/or of auxiliaries in *wh*-questions and in main clause *yes/no*-questions
[61]* vs. [62]*	prepositional dative vs. double object structures after the verb TO GIVE

References

M. S. Aldenderfer and R. K. Blashfield (1984), *Cluster Analysis*, Quantitative Applications in the Social Sciences (Newbury Park, London, New Delhi).

B. Alewijnse, J. Nerbonne, L. van der Veen, and F. Manni (2007), 'A Computational Analysis of Gabon Varieties', in P. Osenova, ed., *Proceedings of the RANLP Workshop on Computational Phonology*. 3–12.

L. Anderwald (2009), *The Morphology of English Dialects* (Cambridge).

D. Biber, S. Conrad, and R. Reppen (1998), *Corpus Linguistics: Investigating Language Structure and Use* (Cambridge).

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

H. Goebl (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* (Wien), 133–172.

W. Heeringa (2004), *Measuring dialect pronunciation differences using Levenshtein distance* (Ph. D. thesis, University of Groningen).

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

C. Hoppenbrouwers and G. Hoppenbrouwers (2001), *De indeling van de Nederlandse streektalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM* (Assen).

F. Inoue (1996), 'Subjective Dialect Division in Great Britain', *American Speech*, 71(2), 142–161.

J. B. Kruskal and M. Wish (1978), *Multidimensional Scaling*, Volume 11 of *Quantitative Applications in the Social Sciences* (Newbury Park, London, New Delhi).

W. Labov (1966), 'The linguistic variable as a structural unit', *Washington Linguistics Review*, 3, 4–22.

A. McMahon, P. Heggarty, R. McMahon, and W. Maguire (2007), 'The sound patterns of Englishes: representing phonetic similarity', *English Language and Linguistics*, 11(1), 113–142.

J. Nerbonne, 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* (Berlin), 647–654.

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

M. R. Spruit (2005), 'Classifying Dutch dialects using a syntactic measure: the perceptual Daan and Blok dialect map revisited', *Linguistics in the Netherlands*, 22(1), 179–190.

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

M. R. Spruit (2008), *Quantitative perspectives on syntactic variation in Dutch dialects* (Ph. D. thesis, University of Amsterdam).

M. R. Spruit, W. Heeringa, and J. Nerbonne (to appear), 'Associations among Linguistic Levels', *Lingua*.

B. Szmrecsanyi (forthcoming), *Woods, trees, and morphosyntactic distances: traditional British dialects in a corpus-based dialectometrical view* .

B. Szmrecsanyi and N. Hernández (2007), *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S")*. `http://www.freidok.uni-freiburg.de/volltexte/2859/` (Freiburg).

M. Wieling, W. Heeringa, and J. Nerbonne (2007), 'An aggregate analysis of pronunciation in the Goeman-Taeldeman-van Reenen-Project data', *Taal en Tongval*, 59(1), 84–116.

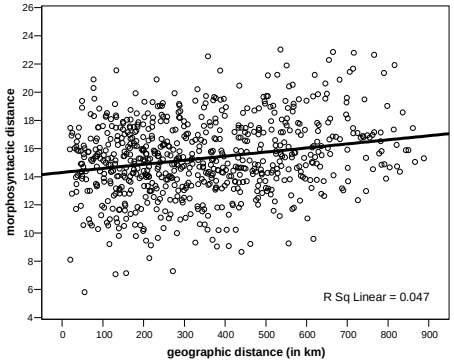

Figure 1: Correlating linguistic and geographic distances, county level ($N = 38$), all features ($p_{total} = 62$), $r = .22$, $p = .00$.

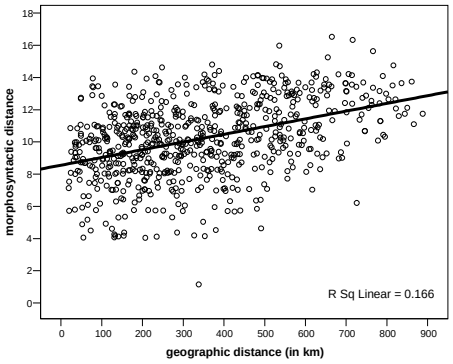

Figure 2: Correlating linguistic and geographic distances, county level ($N = 38$), geographically significant features only ($p_{geo} = 23$), $r = .41$, $p = .00$.

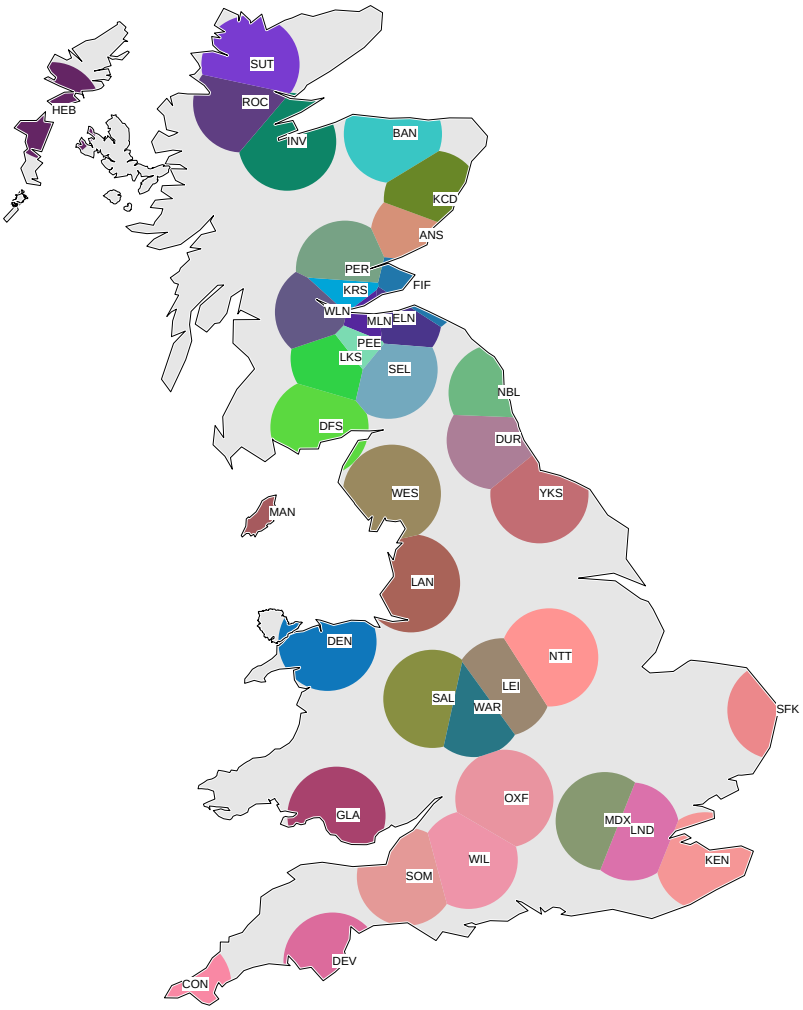

Figure 3: Continuum map: regular MDS on Euclidean distance matrix (county level). Labels are three-letter Chapman county codes (see `http://www.genuki.org.uk/big/Regions/Codes.html` for a legend). Smooth colour transitions indicate the presence of a dialect continuum. Reddish colours correlate best with increased frequencies of multiple negation, greenish colours correlate best with higher frequencies of non-standard weak past tense and past participle forms, and bluish colours correlate best with increased frequencies of *wh*-relativisation.

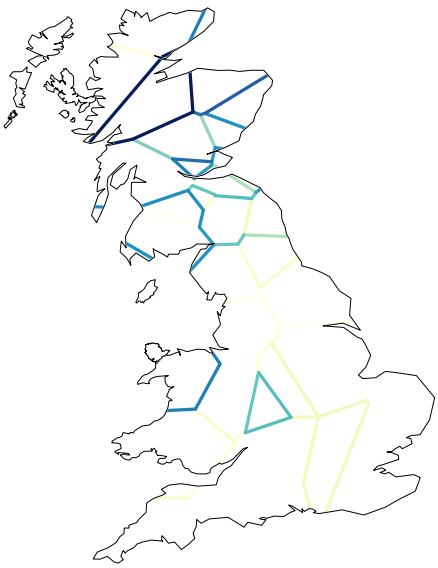

Figure 4: Composite cluster map, county level ($N = 38$), all features ($p_{total} = 62$); input: cophenetic distance matrix (clustering algorithm: WPGMA). Darker borders indicate more robust dialect boundaries.

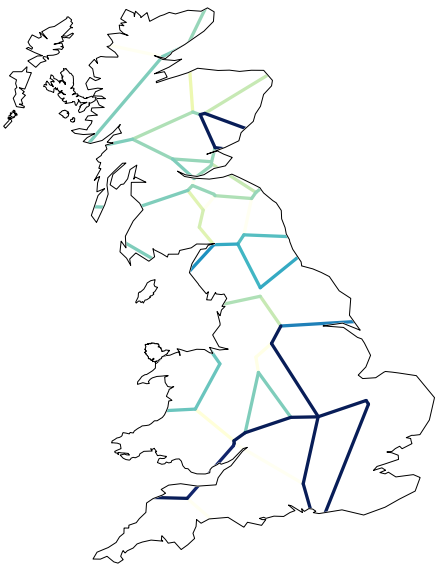

Figure 5: Composite cluster map, county level ($N = 38$), all features ($p_{total} = 62$); input: cophenetic distance matrix (clustering algorithm: WARD). Darker borders indicate more robust dialect boundaries.

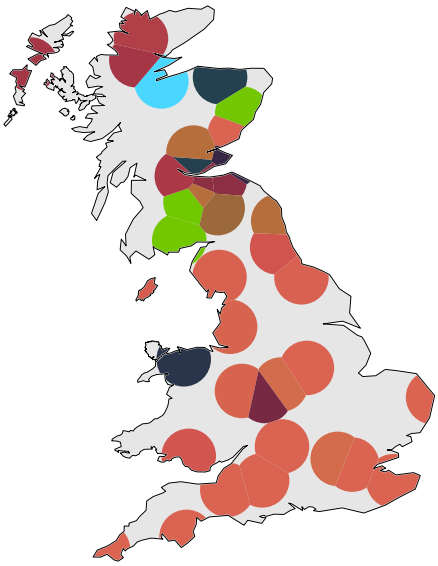

Figure 6: Fuzzy MDS map, county level ($N = 38$), all features ($p_{total} = 62$); input: cophenetic distance matrix (clustering algorithm: WPGMA); felicitousness of the MDS solution: $r = .96$. Similar colours indicate likely membership in the same dialect area.

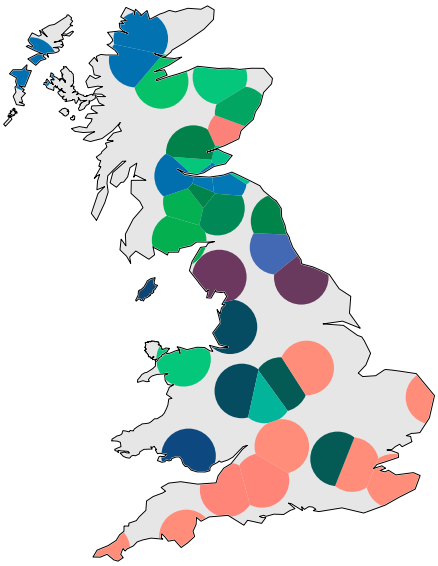

Figure 7: Fuzzy MDS map, county level ($N = 38$), all features ($p_{total} = 62$); input: cophenetic distance matrix (clustering algorithm: WARD); felicitousness of the MDS solution: $r = 1.00$. Similar colours indicate likely membership in the same dialect area.