Forests, trees, corpora, and dialect grammars

Benedikt Szmrecsanyi (University of Leuven)

Abstract

The paper explores how and to what extent morphosyntactic variability in 34 traditional British English dialects is structured geographically. Taking an interest in the forests rather than individual trees, the study presents an aggregate analysis of text frequencies of dozens of morphosyntactic features in a major naturalistic dialect corpus. I sketch the basics of the method, present ways to depict aggregate variability visually, and I also discuss why aggregate analyses based on graded text frequencies are probably more realistic than, e.g., atlas-based aggregation projects subject to high levels of data reduction.

1. Introduction

This contribution sketches ways to milk naturalistic speech data for a geolinguistic signal by investigating joint frequency variance of many linguistic features, a line of analysis that has been referred to as corpus-based dialectometry elsewhere (Szmrecsanyi 2008, 2011, 2012, 2013; Szmrecsanyi and Wolk 2011). The objective of the research discussed here is to explore the relationship between dialect distances and geography, for the sake of enhancing our understanding of how and to what extent language variation is a function of language-external factors, such as geographic space or, implicitly, the likelihood of social contact. Along the way, we will also address some interesting methodological issues.

What is dialectometry? Whereas old-school dialectologists typically study “trees”, (that is, individual dialect phenomena, one feature at a time), dialectometrical inquiry (Goebel 1982; Nerbonne, Heeringa, and Kleiweg 1999; for seminal work, see Séguy 1971) is interested in the “forests” and endeavors to identify “general, seemingly hidden structures from a larger amount of features” (Goebel and Schiltz 1997: 13). This means that dialectometricians put a strong emphasis on quantification, cartographic visualization, and exploratory data analysis to infer patterns from feature aggregates. Empirically, the bulk of the orthodox dialectometry literature in the Séguy-Goebl-Nerbonne tradition draws on linguistic atlas material as its primary data source. For example, Goebel (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, Heeringa, and Kleiweg (1999) analyze aggregate accent distances between 104 Dutch and North Belgian dialects on the basis of 100 word transcriptions provided in the Reeks Nederlands(ch)e Dialectatlassen (RND) (see Heeringa 2004 for a more extensive dialectometrical study of the RND).

Now, while dialect atlases are beautiful and a truly indispensable tool in dialectology, I shall argue in this paper that naturalistic speech corpora are a complementary data source that can possibly provide a more realistic picture of the relationship between dialect variability and geography. To furnish a case study, I turn to the Freiburg Corpus of English Dialects (Hernández 2006; Szmrecsanyi and Hernández 2007), a naturalistic speech corpus that 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

---

1 Sure enough, there are some counterexamples (for example, Bolognesi and Heeringa 2002, 2005), but on the whole it is fair to say that Séguy-Goebl-Nerbonne work is typically atlas based.
1990. In most cases, a fieldworker 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 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 34 different pre-1974 counties in Great Britain including the Isle of Man and the Hebrides. To mitigate data sparsity, the level of areal granularity investigated in the present study will be the county level, which leaves us with 34 dialect objects. The study defines a set of 57 morphosyntactic features (essentially, the usual suspects in the literature), measures these features’ text frequencies in the corpus, and subsequently explores the feature portfolio’s joint frequency variability as a function of language-external parameters, such as geography.

This paper is structured as follows. In Section 2, I explain why the corpus-cum-aggregation endeavor is profitable scientifically. The aim of Section 3 is to sketch how to obtain an appropriate dataset. In Section 4, I present a number of ways to depict aggregate dialect relationships visually. Section 5 elucidates why corpus-based measurements may be more realistic than to atlas-based measurements. Section 6 offers some concluding remarks.

2. Why we need corpus-based dialectometry

Combining the qualitative-philological jeweler’s-eye perspective afforded by the corpus linguistic method with the quantitative-aggregational bird’s-eye perspective, which is the hallmark of dialectometry, is desirable for two principal reasons.

First, multidimensional objects, such as dialects, call for aggregate analysis techniques. That rigorous dialectology requires aggregation (in dialectological parlance, bundling) has been known for quite a while. In the 1930s already, Bloomfield argued, with a view to German dialect geography, that

a set of isoglosses running close together in much the same direction – a so-called bundle

of isoglosses – evidences a larger historical process and offers a more suitable basis of
classification than does a single isogloss that represents, perhaps, some unimportant
feature. (Bloomfield 1933: 342)

The point is that in dialectology, so-called “single-feature-based studies” (Nerbonne 2009: 176) are fine when the research question is tree-centered – i.e. when it is the features themselves that are of analytic interest. But single-feature-based studies are inadequate when it comes to characterizing “forestry”, multidimensional objects such as dialects or varieties (or relations between them). Outside linguistics, this sort of inadequacy is well-known: Economists, for instance assess the economic climate not on the basis of individual macroeconomic indicators (e.g. unemployment), but also consider inflation, GDP per capita, interest rates, and so on. The problem with single-feature-based studies, in linguistics and everywhere else, is that feature selection is ultimately arbitrary (Viereck 1985: 94), and that the next feature down the road may or may not contradict the characterization suggested by the previous feature. There is no guarantee that different dialects will exhibit the same distributional behavior in regard to different features, because isoglosses do not necessarily overlap (Bloomfield 1933: 329). Moreover, individual features may have fairly specific quirks to them that are irrelevant to the big picture. This is why “[s]ingle-feature studies risk being overwhelmed by noise, i.e., missing data, exceptions, and conflicting tendencies” (Nerbonne 2009: 193). So, the aggregate perspective – in Goeb’s parlance, “the synthetic interpretation” of linguistic data (Goeb 2006: 415) – is called for when the analyst’s attention is turned to the forest (i.e. the multitude of features that characterize a given dialect), not the trees (i.e. individual features of a dialect). Aggregation mitigates the problem of feature-
specific quirks, irrelevant statistical noise, and the problem of inherently subjective feature selection, and thus provides a better description of dialects, and a more robust linguistic signal. This robust linguistic signal also facilitates *comparison* of different forests (that is, dialects), which is after all a key objective in dialectometry.

Second, *compared to linguistic atlas 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 */bU_s*/”, while corpus-based (that is to say, frequency-based) approaches seek generalizations along the lines of ‘in Nottinghamshire English, multiple negation is twice as frequent (6 occurrences per ten thousand words) in actual speech than in Yorkshire English (3 occurrences per ten thousand words)’. To be sure, the atlas-based method has many advantages, in particular a fairly widespread availability of data sources and superb areal coverage. Dialect corpora are less wide-spread, and their areal coverage is typically inferior to that of dialect atlases. Nonetheless, as a data source, corpora offer two exciting advantages over dialect atlases. First and foremost, the signal provided by conventional dialect atlases is categorical, exhibits a high level of data reduction, and may hence be less accurate than the corpus signal, which can provide graded frequency information. This highlights the most crucial difference between atlas-based and corpus-based dialectometry: *corpus-based dialectometry is frequency-based dialectometry in its purest form.*

Although the exact cognitive status of text frequencies is admittedly still unclear (for example, we do not know about the precise extent to which corpus frequencies correlate with psychological entrenchment; see Blumenthal 2011), it seems plausible to assume that text frequencies better match the perceptual reality of linguistic input than discrete atlas classifications do. Second, the atlas signal is non-naturalistic and meta-linguistic in nature, and is more often than not concerned with linguistic knowledge. It typically relies on elicitation and questionnaires, and is analytically twice removed, via fieldworkers and atlas compilers, from the analyst. By contrast, text corpora provide more direct, usage-based access to language form and function, and may thus yield a more realistic and trustworthy picture (Chafe 1992; Leech, Francis, and Xu 1994). The well-known major intrinsic drawback of the corpus-based method is that it does not easily address rare phenomena. But then again, one may wonder 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. How to obtain an appropriate dataset

The first step towards corpus-based aggregate analysis consists of defining the feature catalogue. The analyst should aim to base the analysis on as many linguistic features as possible. In the case study at hand, the relevant literature on morphosyntactic variability in varieties of English was surveyed, and suitable dialect phenomena were identified. This resulted in a list of $p = 57$ features falling into eleven major grammatical domains. This list overlaps with but is not identical to comparative morphosyntax surveys in the spirit of Kortmann and Szmrecsanyi (2004) and Szmrecsanyi and Kortmann (2009a, 2009b). The list of course also draws on the battery of grammatical features tested in the *Survey of English Dialects* (Orton and Dieth 1962). A detailed discussion of the features in the catalogue is beyond the scope of this contribution, but the Appendix provides the complete list of features.

The second step involves measuring feature frequencies and creating a so-called *frequency matrix*. In terms of our case study, I would like to refer the reader to Szmrecsanyi

---

2 The present study’s methodology is somewhat similar to the pioneering, frequency-based dialectometry approach of C. Hoppenbrouwers and G. Hoppenbrouwers (1988, 2001). See Heeringa (2004: 16–20) for a discussion in English.
(2010), who provides detailed coding schemes and discusses the technicalities of the extraction process in ample detail. It should be emphasized, though, that the creation of the dataset involved very substantial hand-coding (specifically, > 85000 manual coding decisions), and took into account very meticulously the contexts in which individual features occur. This is why the study reported here does not merely rely on the law of large numbers. First and foremost, it relies on philological and dialectological acumen to identify dialect features in interactional discourse.

After frequency normalization and an optional log-transformation to de-emphasize large frequency differentials, we 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 feature frequency in a particular dialect. Our case study thus yields a $34 \times 57$ frequency matrix: 34 British English dialects, each characterized by a vector of 57 text frequencies. To illustrate: in the dialect corpus, the county Cornwall has a textual coverage of 12 interviews totaling about 107000 words of running text (interviewer utterances excluded). In this material, feature [34] (negative contraction, e.g. they won’t do anything) occurs 326 times, which translates into a normalized text frequency of $326 \times 10000 / 107000 \approx 30$ occurrences per ten thousand words. This figure in turn yields a log-transformed frequency value of $\log_{10}(30) \approx 1.5$, which gives us the frequency that characterizes the measuring point Cornwall in regard to feature [34].

Frequency matrices can serve as input to some aggregational analysis techniques, such as Principal Component analysis or Factor Analysis (see Grieve, this volume). Other, more radically aggregational procedures – and these will take center stage in the subsequent section – are empirically based on so-called distance matrices, which are the outcome of transforming an $N \times p$ frequency matrix into an $N \times N$ dissimilarity table. This transformation abstracts away from individual feature frequencies and specifies pairwise distances between the dialect objects considered (similar to distance tables to be found in, e.g., road atlases). How do we calculate aggregate distances? There is a daunting variety of distance measures, yet given the continuous nature of corpus-derived frequency vectors, I advocate usage of the well-known and fairly straightforward Euclidean distance measure (Aldenderfer and Blashfield 1984: 25), which defines the distance between two dialect objects as the square root of the sum of all $p$ squared frequency differentials (see Szmrecsanyi 2011, 2013 for more detail).

4. Visually depicting aggregate dialect relationships

It is fair to say that the sorts of dataset(s) discussed in the previous section are fairly complex. For example, distance matrices characterize each dialect object in regard to its (numerical!) relation to the other $N - 1$ (in our case, 33) dialect objects in the sample. High dimensionalities like these are hard to grasp when one merely eyeballs the data, so what is needed to see the wood for the trees are sparkling visualization techniques, optionally coupled with advanced statistical dimension-reduction machinery. It is such techniques that will take center stage in this section.

---

3 It should be emphasized here that the frequency matrix used in the present paper specifies absolute feature frequencies. The alternative is a more compact and more genuinely variationist – in the Labovian sense – relative frequency matrix. See Szmrecsanyi (2012: Chapter 2) for a discussion of the differences, and Szmrecsanyi (2008) for a corpus-based dialectometry analysis of a relative frequency matrix.

4 As for reliability, we note for the sake of completeness that the $34 \times 57$ frequency matrix yields a Cronbach’s $\alpha$ value of .77 (see Grieve, this volume, for discussion), a score that comfortably passes the conventional threshold.
Map 1. Beam map. Neighbors that are close morphosyntactically are connected by warm and heavy beams.

Map 2. A similarity map for Nottinghamshire (in white; measuring point 20, English Midlands). Warm tones indicate relative morphosyntactic similarities.

Map 3. A similarity map for Standard British English. Warmer (i.e. reddish) tones indicate relative morphosyntactic similarities. Colder (i.e. bluish) tones indicate relative morphosyntactic dissimilarities.

Map 4. A synopsis map combining the continuum scenario and the dialect area scenario. Clustering algorithm: Ward’s method. Shared variance between MDS solution and original distances: $R^2 = 95.9\%$. Outliers are mapped as language islands.
4.1. Beam maps

In Map 1, we find a Goebl-style beam map that projects our case study’s 34 × 34 distance matrix to geography without much statistical ado. The map restricts attention to so-called “interpoint” (i.e. neighbor) relationships (Goebl 1982: 51). The projection connects morphosyntactically distant neighbors by thin beams, whereas neighbors that are close morphosyntactically are connected by heavy and warm (i.e. reddish) beams. Interpretationally, then, Map 1 highlights a couple of hotbeds of neighborly similarity in Great Britain: in the Southwest of England, for example, there is a comparatively marked axis of interpoint similarities running from Cornwall via Devon and Somerset all the way to Wiltshire. In the Southeast of England, there is a triangle of morphosyntactic similarities connecting Kent, London, and Suffolk. In the Northern Midlands and the North of England, we find a web of strong interpoint similarities encompassing Nottinghamshire, Lancashire, Westmorland, Yorkshire, and Durham. The Central Scottish Lowlands exhibit a bolt of interpoint similarities involving parts of the urbanized “Central Belt”.

4.2. Similarity maps

Map 2 features a Goebl-style similarity map for the arbitrarily chosen measuring point Nottinghamshire. In similarity maps, dialect sites are colored according to their similarity to some particular measuring point (in this case Nottinghamshire), which appears in white. As with beam maps, warmer colors indicate relative similarities, while colder shades depict relative dissimilarities (see, for example, Goebl 2007: 140). According to Map 2, then, Nottinghamshire entertains particularly close linguistic relationships to many other counties in England, while it is dissimilar morphosyntactically from most measuring points in Scotland. Map 2 thus clearly exhibits a geographic signal.

Map 3 is also a similarity map, but unlike in Map 2 the reference point is not a dialectal measuring point but Standard British English. This in a geolinguistic perspective “artificial” measuring point is furnished by extracting feature frequencies not from the dialect corpus FRED but from a small reference corpus sampling extracts from the British component of the International Corpus of English (see Szmrecsanyi 2013 for details). So in Map 3, a color slide distinguishes between dialectal measuring points that are extremely similar and dialectal measuring points that are extremely dissimilar to Standard English; more bluish hues indicate dissimilarities while more reddish hues indicate similarities. The pattern emerging from Map 3 is that the Scottish Lowlands (except for Midlothian and East Lothian) as well as the English Midlands (in particular, Nottinghamshire and Leicestershire) exhibit the largest distances to Standard British English. Conversely, the relatively young dialects in the Scottish Highlands and on the Hebrides are overall quite close, and so are measuring points in the North of England.

4.3. Cluster maps and continuum maps

Dialectological theory offers two ways to think about dialect geographies (Heeringa and Nerbonne 2001): The dialect area scenario, and the dialect continuum scenario. The

---

5 On a technical note, Map 3 was generated by dividing the observable distance range into 10 percentile bins such that each bin contains roughly the same number of observations. Each county's bin rank was then mapped on the red-green-blue color scheme, assigning a perfect blue hue to the highest bin rank, a perfect red tone to the lowest bin rank, and gradient red-blue color blends to the ranks in between. Also note that unlike in Map 1 and Map 2, the radius of the polygons in Map 3 and Map 4 is limited to approximately 60km in order to do more visual justice to the areal coverage of the dialect corpus.
dialect area scenario partitions a given dialect landscape into internally homogeneous but mutually heterogeneous dialect areas. The dialect continuum scenario posits that there are no sharp boundaries between dialects. Instead, linguistic distance is supposed to be directly proportional to geographic distance. The two scenarios have a long history of thought in dialectology (going back to luminaries such as Schleicher 1863; Schmidt 1872). In their extreme manifestations, the two scenarios are not quite compatible. But modern dialectologists believe that reality actually comes in shades of grey (see Goebel 1983), which is why we will assume here that the way in which geographic space conditions dialect variability has a continuum-like structure itself. At one end of this continuum, we find perfect dialect continua where linguistic transitions are maximally smooth; at the opposite end, we observe perfectly regionalized dialect landscapes in which transitions between internally homogeneous dialect areas are maximally abrupt.

How can we approach this issue cartographically? Cluster maps address the dialect area scenario and project the outcome of cluster analysis to geography (see, for example, Goebel 2007). The map type aims to explore the extent to which dialect landscapes are organized along the lines of internally homogeneous but mutually more or less sharply demarcated dialect areas. By contrast, continuum maps, a signature visualization technique developed in Groningen (Nerbonne, Heeringa, and Kleiweg 1999), are primarily concerned with the dialect continuum scenario. The idea is to utilize Multidimensional Scaling (MDS) to assign every dialect polygon on the map a distinct color hue. Smooth color transitions between dialect polygons emphasize the continuum-like nature of the dialect landscape, while abrupt color transitions point to the necessity of alternative explanations.

In Map 4 we find a synopsis map that seeks to marry the two map types, and thus combines the dialect area view and the dialect continuum view in one and the same projection. A synopsis map can created as follows. One initially applies repeated clustering with noise (i = 10000, c = σ, clustering algorithm: Ward’s method) to the original distance matrix (on clustering with noise, see Nerbonne et al. 2008). Next, the resulting cophenetic distance matrix is subjected to MDS. Finally, the artificial MDS coordinates we thus obtain for the first three MDS dimensions are mapped onto the red-green-blue color scheme, to assign each dialect polygon on the map distinct color hues. Where is the interpretational meat? Compared to a genuine continuum map, Map 4 accentuates differences between dialects that belong to different, inductively identified clusters or dialect groupings. But compared to a genuine cluster map, Map 4 emphasizes the gradient nature of dialect grouping memberships. Against this backdrop, the picture that emerges from Map 4 essentially boils down to a tripartite dialect division: we find an orange group of dialects mainly in the South of England; a violet-bluish group of dialects located mainly in the North of England (plus measuring points in the Scottish Highlands); and a greenish cluster of Scottish Lowlands dialects. Note, however, that the orange and violet-bluish England-centered groups are fairly homogeneous internally: here, we do not find a lot of internal color variance (read: a low degree of morphosyntactic variability), probably thanks to a long-term process of region-internal dialect leveling. By contrast, the greenish Scottish Lowlands group exhibits a good deal of internal gradience. In point of fact, the greenish group emerges as a fairly well-behaved dialect continuum internally. It seems, therefore, as though England’s internal morphosyntactic dialect landscape is organized along the lines of a dialect area scenario,

Map 4 represents a number of outlier dialects as language islands – small circles – so as not to disturb the overall picture.

I acknowledge that some authors have objections to using Ward’s method as a clustering algorithm; see Heeringa (2004: 150–153) for discussion.
while Central Scotland’s morphosyntactic dialect landscape obeys the laws of a dialect continuum.

4.4. Network diagrams

Figure 1 displays a *NeighborNet network diagram* (Bryant and Moulton 2004).\(^8\) Originally developed to address evolutionary biology issues, network diagrams and related visualization techniques are increasingly popular to gauge genetic relationships and aggregate similarities between languages (e.g. Dunn et al. 2005) and language varieties and dialects (e.g. A. McMahon et al. 2007). Their big advantage vis-à-vis simple cluster analysis (see previous section) is that network diagrams are able to detect and depict conflicting signals, and that they can represent the effects of language contact:

When interpreting such networks, the equivalents of edges connecting two tree nodes in a dendrogram are either individual lines, or sets of parallel lines. In this network, we only find individual lines directly at the leaf nodes, and many sets of parallel lines, combining to the boxy shapes that form the body of the network. Each represents a way of splitting the total set of dialects into exactly two groups. The longer a given line or set of lines, the greater the difference between the groups. To give an example, the comparatively large vertical set of lines directly below the point where Durham joins the network divides the dialects into the following two groups: one group that consists of Nottinghamshire as well as all Southwestern and Southeastern dialects except Middlesex, and another group that contains all other dialects. When two such divisions are not representable as strictly hierarchical, the resulting lines form boxy shapes. (Szmrecsanyi and Wolk 2011: 575–576)

\(^8\) The diagram and its interpretation is reproduced from Szmrecsanyi and Wolk (2011).
Figure 1. Network representation of morphosyntactic distances. Symbols indicate *a priori* dialect areas.

The correspondence in Figure 1 between geographical location and placement in the network diagram is certainly not perfect, but once again there clearly is a geographical signal. Most Southern dialects can be found at the bottom of the diagram. We then “move up”, clockwise, via the English Midlands and Northern English dialects toward the Scottish dialects at the top of the diagram. Upon closer inspection, it turns out that while some of the large-scale dialect areas, such as the (mostly) Southern group discussed in the previous section, are actually represented by an individual split, others – such as the North of England group – are not actually a single group, but rather a collection of smaller ones with interlocking resemblances.

4.5. Cartography and diagrams: interim summary

The various diagrams and projections to geography presented in this section would seem to have suggested that in the aggregate perspective, the text frequency portfolio extracted from the dialect corpus FRED clearly exhibits a geographic signal. That said, though, we note that the signal is not always geographically consistent, dialect areas are not perfectly coherent, and so on – in short: the data source is noisy. It is precisely because of this that dialect-atlas-derived visualizations are much “neater”, as a casual perusal of the atlas-
based dialectometry literature suggests. The task before us in the next section is to determine whether this noisiness is a flaw, or whether it reflects some truth that is actually rooted in linguistic reality.

5. Why corpus-based analysis may be more realistic than atlas-based analysis

The name of the game in this section is number crunching. We will begin by regressing morphosyntactic dialect distances against a number of language-external, geography-related distance measures for the sake of calculating the exact strength of the geolinguistic signal in the corpus material. Subsequently, we will turn to the atlas-based literature to compare signal strengths. Addressing the discrepancies that this comparison yields, we will manipulate our corpus-based dataset in a number of ways to explore the reasons for the discrepancies.

To set the scene, let us correlate our morphosyntactic distance matrix with three language-external distance matrices:

*as-the-crow-flies distances.* It is computationally trivial to calculate pairwise as-the-crow-flies distances. 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; Nerbonne et al. 1996; Shackleton 2007).

*least-cost travel times.* It is plausible that what really matters in terms of dialect distances is not what crows do but how long it would take a human traveler to get from point A to point B (Gooskens 2005; Szmrecsanyi 2012). To calculate this measure, I turned to Google Maps (http://maps.google.co.uk/), 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. I queried Google Maps for all $34 \times 33 / 2 = 561$ unique dialect pairings in the $34 \times 34$ distance matrix, thus obtaining pairwise least-cost-travel time estimates.

*linguistic gravity indices.* Sociolinguist Peter Trudgill has 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” (1974: 233). Using Trudgill’s formula, I calculated linguistic gravity values for each of the 561 dialect pairings in our database, feeding in least-cost travel time as geographic distance measure and early twentieth century population figures$^9$ (in thousands) as a proxy for speaker community size.

---

$^9$ 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 2. Morphosyntactic dialect distances versus (a) as-the-crow-flies distances \((r = .21, \ p < .001, \ R^2 = 4.4\%)\), (b) least-cost travel times \((r = .27, \ p < .001, \ R^2 = 7.4\%)\), and (c) Trudgill’s linguistic gravity indices (log-transformed) \((r = −.49, \ p < .001, \ R^2 = 24.1\%; \ logarithmic \ estimate)\). Each dot represents one of \(N = 561\) unique dialect pairings. Solid lines are regression lines.

Figure 2 provides three scatterplots that plot 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 the figure caption suggest that as-the-crow flies distance accounts for 4.4% of the morphosyntactic variance, least-cost travel time for 7.4%, and linguistic gravity for 24.1%. 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.

While 24.1% is a figure one can start writing home about, I note that in comparison to figures reported in the atlas-based dialectometry literature this \(R^2\) value is quite low. For example, Shackleton (2007), in his study of phonetic variation in the Survey of English Dialects, reports \(R^2\) values of up to 66% 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% for the relation between syntax and geography. I emphasize that the impressive \(R^2\) values reported by Shackleton and Spruit et al. draw on the most unsophisticated geographic distance predictor – as-the-crow-flies distance – which in our corpus-based dataset explains no more than 4.4% of the variance in linguistic distances. This discrepancy deserves some discussion, and in what follows I will sketch three possible explanations.

5.1. Great Britain is different

This explanation argues that the dialect landscape in Great Britain is simply organized differently, with less reference to geography, than elsewhere. Now, it is true that the presence
of two dialect kernels and the ensuing contrast between English English and Scottish English dialects renders linguistic variability in Great Britain a rather special case. Yet there is no \textit{a priori} reason to assume that a bi-kernel situation like this should necessarily depress the explanatory potential of geography. What is more, the Great-Britain-is-different account fails to explain why, according to the present analysis, morphosyntactic variability in Great Britain exhibits less than five per cent of shared variance with geography whereas Shackleton (2007), who studies atlas-derived phonetic variability in England, obtains an $R^2$ value of 49%.

5.2. Morphosyntax is different

Some scholars would contend that (morpho)syntactic variability in general is different from accent or lexical variability. This argument has more merit, given a long-held suspicion by typologists, syntacticians, and even dialectologists (for example, Lass 2004; Löffler 2003; Wolfram and Schilling-Estes 1998) that morphosyntax and grammar are less diffusable geographically than, for instance, pronunciation. But there are some problems with this assumption. On empirical grounds the analysis of aggregate syntactic differences in the Dutch language area presented in Spruit et al. (2009) does yield a substantial overlap between syntactic and geographic distances, and so plays havoc with the putative non-diffusability of syntax. Also, a perusal of the literature calls into question the axiom that grammar is hostile to geographic diffusion. Yes, according to the famous borrowing scale in Thomason and Kaufman (1988) structural features are not borrowed as easily as low-level lexical and maybe phonetic features. However, Heine and Kuteva (2005: 1) point out how the alleged fact that “language structure is fairly resistant to change in situations of language contact” is no more than a prejudice that ultimately goes back all the way to Ferdinand Saussure and Edward Sapir. Against this backdrop – and certain conceptual differences between contact-induced change and diffusion notwithstanding (see Wiemer and Wälchli 2012 for discussion) – I would like to argue that by and large what is true for language contact must also be true for dialect contact, and hence there is no good reason why morphosyntax should not diffuse geographically.
Figure 3. Morphosyntactic dialect distances (selected feature sets) versus as-the-crow-flies distances. (a): all \( N = 34 \) measuring points, all 57 (linear estimate: \( R^2 = 4.4\%, \ p < .001 \)). (b): all \( N = 34 \) measuring points, 14 morphosyntactic features which are geographically significant on an individual basis (linear estimate: \( R^2 = 14.6\%, \ p < .001 \)). Solid lines depict linear regression estimates.

5.3. Atlas-based measurements are not accurate

The most convincing explanation is all about methodology: compared to corpus-based and frequency-centered approaches, atlas-based approaches overestimate the importance of geography. First, I stress that feature selection does matter a great deal. To put this into sharper empirical focus, Figure 3 displays two scatterplots which visually depict the relationship between pairwise morphosyntactic distances (on the vertical axes) and pairwise as-the-crow-flies distances (on the horizontal axis). The left plot, which we had already seen in Figure 2, depicts morphosyntactic distances on the basis of frequencies of all 57 features on which the present study’s aggregation endeavor is based. The right plot’s morphosyntactic distance calculation, however, restricts attention to those 14 morphosyntactic features in the feature catalogue which have, on an individual basis, a statistically significant geographic distribution (see Szmrecsanyi 2013 for more detail) – it turns out, for example, that feature [33] (multiple negation, as in don’t you make no damn mistake) is one of the most areal features in the dataset. This, then, is one of the features that feeds into Figure 3b. With Figure 3b depicting a stronger correlation (notice the steeper regression line) because it focuses on the most areal features in the dataset, the point I am driving at is that drawing on specific feature subsets comprising material that is somehow more relevant geographically inevitably boosts the explanatory power of geography – in our case study, from \( R^2 = 4.4\% \) to \( R^2 = 14.6\% \). So it is fair to ask to what extent compilers of linguistic atlases really draw on all available features (as mandated by the dialectometric creed), instead of just those features that seem geographically “interesting” in some way (in which case a comparatively strong showing of geography is to be expected).

In this connection, a reviewer noted that my above empirical argument (“The selection of atlas features is probably geographically biased: if we take a set of geographically biased features, we obtain a better correlation with geography. Hence, atlas data are geographically biased”) does not qualify as unassailable evidence that atlas data are indeed biased. This objection is probably correct. But even so it is, at a minimum, plausible to conjecture that
dialect atlases are biased towards geographically distributed features, and that this bias has ramifications for atlas-based dialectology research.

![Graphs showing morphosyntactic dialect distances versus as-the-crow-flies distances](image)

**Figure 4.** Morphosyntactic dialect distances (various levels of data reduction) versus as-the-crow-flies distances. (a): frequencies, all measuring points, all features (linear estimate: $R^2 = 4.4\%, p < .001$). (b): rank by frequency, all measuring points, all features (linear estimate: $R^2 = 7.1\%, p < .001$). (c): two frequency bins, all measuring points, all features (linear estimate: $R^2 = 9.2\%, p < .001$). Solid lines depict linear regression estimates.

Secondly, I suspect also that the more or less violent data reduction that is the hallmark of most survey and atlas projects systematically overrates geography. What is data reduction?

Data reduction is any summarizing or grouping of data where information might be lost. If your water gage measures on three occasions A 1.03 m, B 1.97 m, and C 5.19 m, the data is reduced to various extents if you take down instead A 1 m, B 2 m, C 5 m; or A low, B middle, C high; or A low, B low, C high. (Wälchli 2009: 77)

Wälchli (2009) argues that data reduction may be a problem for typological analysis, for instance when typologies are generated drawing on atlas databases such as the World Atlas of Language Structures. The same is true, I submit, for dialectology research based on dialectological atlas databases. Recall that atlas data are (typically) categorical and rely on a second-hand “attested” vs. “not attested” distinction. This sort of data reduction may exaggerate the explanatory power of geography because linguistic contrasts and distinctions appear more pronounced than they actually are. Consider the plots in Figure 4, which again depict the relationship between as-the-crow-flies geographic distances and morphosyntactic distances in our case study dataset. As in the previous figure, Figure 4a is a reference plot whose morphosyntactic distance measurements are based on maximally graded feature frequencies, of the sort I have used throughout this study. Figure 4b depicts morphosyntactic distances which have been calculated on the basis of a frequency rank measure – for any given morphosyntactic feature, the measuring point with the highest text frequency receives rank 1, the measuring with the lowest frequency receives rank 34, and so on. This is a mild form of data reduction because information about the exact scope of frequency differences between two neighboring ranks is lost. Figure 4c calculates morphosyntactic distances subject to more muscular data reduction: measuring points are categorized, for each individual
morphosyntactic feature, into two frequency bins, low-frequency (below-median) versus high-frequency (above-median). In other words, Figure 4c essentially transforms the dataset's graded frequency signal into a reductionist “high”/“low” signal (akin to a “yes”/“no” signal). Note now that data reduction seems to increase the explanatory power of geography – in Figure 4, from about 4% to about 9% tops. The reason for this boost is that data reduction exaggerates contrasts and removes noise,\(^{10}\) and I suggest that atlas-based dialectometry suffers from this sort of distortion.

5.4. Corpus-based versus atlas-based dialectometry: interim summary

By way of an interim summary, notice that an “intelligent” combination of outlier removal, data reduction, and feature selection will boost the explanatory power of as-the-crow-flies distances in the dataset at hand to an \(R^2\) value of about 25%, a score which is within normal parameters of variation in the atlas-based dialectometry literature. But at what price? The answer boils down to the question what the most appropriate method is for assessing how dialects (and dialect relationships) “actually are”. In other words, is noise necessarily a measuring inaccuracy, or is it – as I have suggested – a “fact” that is there in the real world and that ought to be captured in dialectological analysis? These are, of course, issues that do not have easy answers, and dialectologists are not exactly drowning in research addressing this issue. Perceptual dialectology experiments (in the spirit of e.g. Inoue 1999; Niedzielski and Preston 1999; Preston 1999) may be our best bet for a relatively objective answer. But in the absence of other evidence we should maintain that feature selection and data reduction are essentially a form of academic fraud. Geographically insignificant features and frequency noise are probably part of linguistic reality, and any big-picture account should have to deal with these things, instead of disposing of them for the sake of neatness.

6. Conclusion

This contribution has adopted a frequency-based and forest-centered approach to gauge dialectal grammatical variability. In this spirit, we have explored methodologies to tap naturalistic speech corpora for the sake of modeling aggregate grammatical relationships between dialects as a function of properties of geographic space. In so doing, the study departs from previous work in dialectology because (i) it draws on frequency vectors that derive from naturalistic corpus data, and not on discrete atlas classifications, and because (ii) it marries the careful analysis of dialect phenomena in context to aggregational-dialectometrical analysis techniques.

The first key assumption underpinning this contribution was that holistic, aggregational analysis (the “forest” perspective) is superior to particularistic, single-feature-based analysis (the “tree” perspective) whenever the analyst’s interest lies with dialect relationships per se, and not with individual features, constructions, or functions. The second key assumption was that compared to linguistic atlas material, corpora yield a more realistic linguistic signal, thanks to the graded frequency information that is the hallmark of corpus-linguistic analysis and because of the more direct access to language usage that corpora provide.

On the empirical plane, we have seen that the frequency signal that can be extracted from speech data and depicted graphically by various map and diagram types is clearly

\(^{10}\) It should be added that we cannot entirely rule out that corpus-linguistic feature extraction technologies add some measuring inaccuracies, although I would like to reiterate that much of the dataset subject to analysis here derives from fairly clean hand-coding (Szmrecsanyi 2010).
patterned geolinguistically, although in comparison to atlas-based dialectometrical inquiry it is beset by more noise and a weaker relationship between linguistic and geographic distances. But I have offered that this noisiness is actually a nice thing to have, not a bug: the corpus signal is simply more realistic than the atlas signal, which is subject to data reduction and feature selection. Linguistic reality may not be as neat as atlas-based dialectometry maps might suggest.

The idea, I hasten to add, that corpus-based aggregation profits linguistic analysis is fairly old news to multidimensional register analysts in the tradition of Biber (1988), who have been skillfully pursuing this line of analysis since the 1980s. Dialectologists have been slower to appreciate the promise of corpus-based analysis, but during the past ten years, a number of dialect corpora have been coming online, and more and more corpus-based dialectology analyses – some of them aggregational – are finding their way into the discipline’s publication outlets. Analysts interested in variation across languages admittedly face the most formidable obstacles to the corpus-cum-aggregation endeavor. For one thing, there are very many languages but, at least outside the European context, very few corpora. What is more, the discipline has shifted away in recent years from a “whole language”, holistic approach to a more narrow focus on specific linguistic domains. But be that as it may, I still believe that all those analysts that are in the business of, e.g., “disentangling geography from genealogy” (which is the title of a paper, not included in this volume, presented by Michael Cysouw at the FRIAS workshop), such as Sprachbund theorists (e.g. Lindstedt 2000) or analysts interested in structural and genealogical diversity in space (e.g. Nichols 1992) and/or time (e.g. Nettle 1999), are likely to reap benefits from looking at graded frequencies of many features in corpora of text or speech.

Appendix: 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)

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 knewed 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 3rd person singular subjects (e.g. if this man don’t come up to it)
[41] standard doesn’t with 3rd 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)
absence of auxiliary be in progressive constructions (e.g. I said, How ∅ you doing?)

[44] non-standard was (e.g. three of them was killed)
[45] non-standard were (e.g. he were a young lad)

I. Relativization

[46] wh-relativization (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 ∅ 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])
References


Hoppenbrouwers, Cor and Geer Hoppenbrouwers 2001 *De indeling van de Nederlandse streektaalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Assen: Koninklijke Van Gorcum.


Nerbonne, John, Peter Kleiweg, Franz Manni, and Wilbert Heeringa 2008 Projecting dialect differences to geography: bootstrapping clustering vs. clustering with noise. In:


Schleicher, August 1863 *Die Darwinsche Theorie und die Sprachwissenschaft. Offenes Sendschreiben an Herrn Dr. Ernst Haeckel, o. Professor der Zoologie und Direktor des zoologischen Museums der Universität Jena*. Weimar: Böhlau.

Schmidt, Johannes 1872 *Die Verwantschaftsverhältnisse der indogermanischen Sprachen*. Weimar.


