Benedikt Szmrecsanyi
(Albert-Ludwigs-Universität Freiburg i. Br.)

Geography is overrated[1]

1.   Introduction

Dialectologists and geolinguists tend to instinctively assume that geographic proximity predicts dialectal similarity (or, conversely, that geographic distance predicts dialectal distance). Nerbonne & Kleiweg (2007: 154) have referred to this axiom as the "Fundamental Dialectology Principle" (henceforth: FDP). This contribution endeavors to ask a number of critical questions about the FDP, using traditional British English dialects as an empirical testing site: What is the exact extent to which geographic distance predicts morphosyntactic distance? What is the best way to operationalize 'geographic distance' (as-the-crow-flies distance versus least-cost travel time)? Has the FDP a gradient effect such that there are no abrupt dialect boundaries (this we will refer to as the 'continuum view'), or is the so-called 'dialect area view' more accurate, in that the FDP impacts linguistic variability via geographically more or less coherent dialect areas (Heeringa & Nerbonne 2001)? We shall see that mere geographic distance is a poor predictor of aggregate morphosyntactic variability.

On methodological grounds, the present study marries philologically responsible corpus-based research on morphosyntactic variability in British English dialects to aggregational-dialectometrical analysis techniques. Corpus-based dialectology (Anderwald & Szmrecsanyi 2009) is a methodology that draws on principled collections of naturalistic texts to explore authentic dialect usage, typically with a narrow focus on particular dialect features. Dialectometry (Séguy 1971; Goebl 1982; Nerbonne, Heeringa & Kleiweg 1999) is the branch of geolinguistics concerned with measuring, visualizing, and analyzing aggregate dialect similarities or distances. Crucially, orthodox dialectometry draws on linguistic atlas data (typically on accent differences) as its primary data source. Against this backdrop, our approach (see also Szmrecsanyi 2008, 2011) is original in three ways. It seeks, first, to measure

aggregate distances and similarities between traditional dialects in the British Isles, taking into account the joint variability of dozens of grammatical phenomena. It is thus arguably less arbitrary than studies basing their claims on the distribution of a single dialect feature. Second, we will be concerned with *morphosyntactic* variability, which is underresearched vis-à-vis pronunciational or lexical dialect variability. Third, the present study relies not 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 approaches drawing on *categorical* and thus mediated and reductionist (cf. Wälchli 2009) input data.

2.   Data

We use the *Freiburg English Dialect Corpus* (henceforth: FRED) (Hernández 2006; Szmrecsanyi & Hernández 2007) as the primary data source for our inquiry. 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 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 '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 level of area granularity investigated in the present study will be the county level. From the 38 counties sampled in FRED, we removed four counties (Kinross-shire, Inverness-shire, Fife, and Lanarkshire) with comparatively thin coverage (< 5,000 words of running text).

   Note that FRED 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.

3.   Method

The first task was to define a catalogue of morphosyntactic features which would serve as the basis of our aggregate analyses.
   In keeping true to the spirit of dialectometrical analysis (cf. Goebl 1982; Goebl 1984; Nerbonne to appear), the overarching aim was to include as

many phenomena as possible. To this purpose, we canvassed the dialecto-logical, variationist, and corpus-linguistic literature, and identified suitable phenomena. In this endeavor, particular attention was paid to comparative sources, such as the comparative morphosyntax survey reported in Kortmann and Szmrecsanyi (2004) and the battery of morphosyntactic features covered in the *Survey of English Dialects* (Orton and Dieth 1962). We thus ended up with a list of 57 morphosyntactic features, which are listed in the Appendix, along with linguistic examples.[2] Crucially, for a feature to be included in the catalogue, it did not matter if the feature had previously been reported as geographically distributed 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. The catalogue also contains both fairly categorical and thus somewhat salient non-standard features, which tend to be either present or absent – feature [31] (the negative suffix *–nae*) is again a good example – as well as features whose variation is more statistical in nature, and which are thus arguably less salient. Features [8] and [9] on gradient genitive variation exemplify the latter type. It follows that the features included in the catalogue also differ in terms of their 'standardness' – feature [2] (standard reflexives), for instance, is about perfect standard forms, while feature [28] (non-standard weak verb forms) is not 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 random variability will cancel out in the aggregate view, and that, quoting John Nerbonne, a "large number of variables, even though they will contain a great deal of variation irrelevant to questions of geographic or social conditioning, will nonetheless provide the most accurate picture of the relations among the varieties examined" (Nerbonne 2006: 464). For practical purposes, however (and despite this study's the-more-the-merrier spirit), two criteria had to be met for a candidate feature to be included in the catalogue:

– 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 hits in FRED as a whole.
– The feature also had to be extractable subject to a reasonable input of labor resources by a human coder. This is why, for example, some hard-to-retrieve null phenomena (such as zero relativization) are not considered in the catalogue.

---

[2] This is not the place to discuss the catalogue in any detail, but see Szmrecsanyi (2010) for a detailed discussion of individual features.

The second step was to extract feature frequencies from FRED. 31 features in the catalogue are sufficiently 'surfacy' to be extractable without human intervention. In such cases, retrieval scripts written in the programming language *Perl* did the heavy lifting and established the relevant text frequencies. 26 features in the catalogue required manual disambiguation prior to extraction via *Perl* scripts, and in all the present study's dataset is based on a total of well over 80,000 manual coding decisions (Szmrecsanyi 2010 discusses the technicalities in considerable length). Subsequently, the resulting text frequencies were *log* transformed (a customary procedure to de-emphasize large frequency differentials and to alleviate the effect of frequency outliers) and arranged in a 34 × 57 dimensional frequency matrix (34 counties, each characterized by a vector of 57 text frequencies). As for dataset-internal reliability, this matrix yields a Cronbach's α value (cf. Cronbach 1951) of .77, which is satisfactory for the purposes of the present study.

In a third step, the 34 × 57 frequency matrix was transformed into a 34 × 34 distance matrix, which abstracts away from individual feature frequencies and specifies pairwise distances between the dialects considered. The measure used to calculate these distances was the well-known Euclidean distance measure, where the distance between two dialects is defined as the square root of the sum of all 57 squared frequency differentials.

By way of illustration, Map 1 projects the 34 × 34 Euclidean distance matrix to geography. As a so-called 'beam map' (Goebl 1993: 53), Map 1 restricts attention to interpoint relationships – in other words, only neighboring measuring points are connected, which is a way of distilling the essentials of what distance matrices have to tell us. Morphosyntactically dissimilar neighbors are linked by cold (blueish) and thin beams; neighbors that are close morphosyntactically are connected by warm (reddish) and heavy beams. Visual inspection of the map points to four hotbeds of neighborly similarity in Great Britain:

1. In the Southwest of England, there is a comparatively marked axis of interpoint similarities running from Cornwall via Devon and Somerset all the way to Wiltshire.
2. In the Southeast of England, we note a triangle of relatively modest morphosyntactic similarities connecting Kent, London, and Suffolk.
3. In the Northern Midlands and the North of England, we find a web of strong interpoint similarities encompassing Nottinghamshire, Lancashire, Westmorland, Yorkshire, and Durham.
4. The Central Scottish Lowlands exhibit a bolt of interpoint similarities involving parts of the urbanized 'Central Belt': West Lothian, Midlothian, and East Lothian.

4.   Explaining morphosyntactic distances

The beam map in Map 1 is certainly informative, but it cannot directly answer our questions about the validity of the FDP. To do that, we need to systematically correlate morphosyntactic distances with geography-related language-external variables. It is to this task that we turn next.

4.1. As-the-crow-flies distance

We begin by studying the correlation between morphosyntactic distances and as-the-crow-flies geographic distance, which is the most common geographic distance measure in previous dialectological and dialectometrical study. Using a trigonometry formula on the FRED county coordinates, it is easy to calculate pairwise geographic distances, which can then be correlated against pairwise morphosyntactic distances.

   Figure 1 is a diagram that plots morphosyntactic distances on the vertical axis against as-the-crow-flies distances on the horizontal axis. Every dot in this diagram represents one of the $34 \times 33/2 = 561$ county/county pairings under analysis here. The heavy line is a non-parametric regression line (more
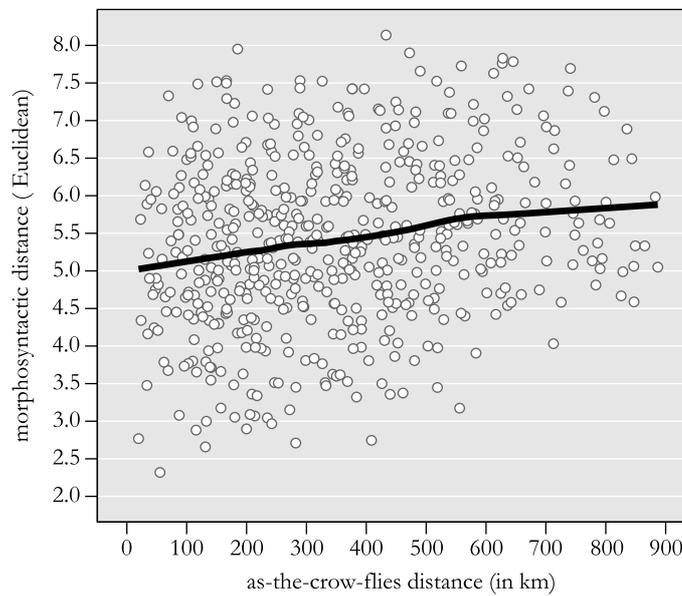

Figure 1: Scatterplot – morphosyntactic distances (vertical axis) vs. as-the-crow-flies distances (horizontal axis). Solid line: non-parametric regression curve. Explained variance (linear estimate): $R^2 = .044$ ($p < .001$).

specifically, a locally weighted scatterplot smoothing curve) that seeks to visualize the overall relationship. As can be seen, there is some sort of relationship here, in that increased as-the-crow-flies distance between two dialects seems to predict increased morphosyntactic distance. However, while significant ($p < .001$), the relationship is quite weak: regression analysis shows that assuming a linear relationship, as-the-crow-flies distance only explains about 4.4% of the variance in morphosyntactic distances ($R2 = .044$). This is a rather low proportion, both in absolute terms and against the backdrop of previous dialectometrical research: Shackleton (2007) investigates atlas material (*viz.* the *Survey of English Dialects*) and reports $R2$ values of up to .66 for the relationship between phonetic and geographic distances in England; Spruit, Heeringa & Nerbonne (2009), in an atlas-based study on aggregate linguistic distances in Dutch dialects, calculate $R2$ values of .47 for the correlation between geography and pronunciation, .33 for lexis, and .45 for syntax. We also note that the relationship in our data between morphosyntactic distances and as-the-crow-flies distances is almost perfectly linear, and not sublinear, as one would expect given Séguy (1971) and much of the subsequent atlas-based dialectometry literature: a logarithmic estimate of the relationship yields an actually significantly ($p < .001$) worse $R2$ value of .041.

By way of an interim summary, we note that there is a linear and comparatively weak relationship between morphosyntactic distances and as-the-crow-flies distances.

4.2. Travel cost

As-the-crow-flies distance between dialect localities is computationally trivial to calculate, but the problem is that speakers do not have wings. Surely, what matters for dialect distances is not what crows do, but how much time it would take a human traveler to get from point A to point B? We thus turned to Google maps (http://maps.google.co.uk/), which has a router finder facility 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. Least-cost travel time was thus established for all 561 county/county pairings subject to analysis here, which resulted in a least-cost travel time matrix.[3] We acknowledge that matching linguistic data sourced

---

[3] The travel times to be analyzed here assume travel by car. We experimented with Google's 'walking' option, but found that this parameter yields a substantially lower correlation with linguistic distances. Similarly, least-cost travel distance appears to be a measure inferior to least-cost travel time.
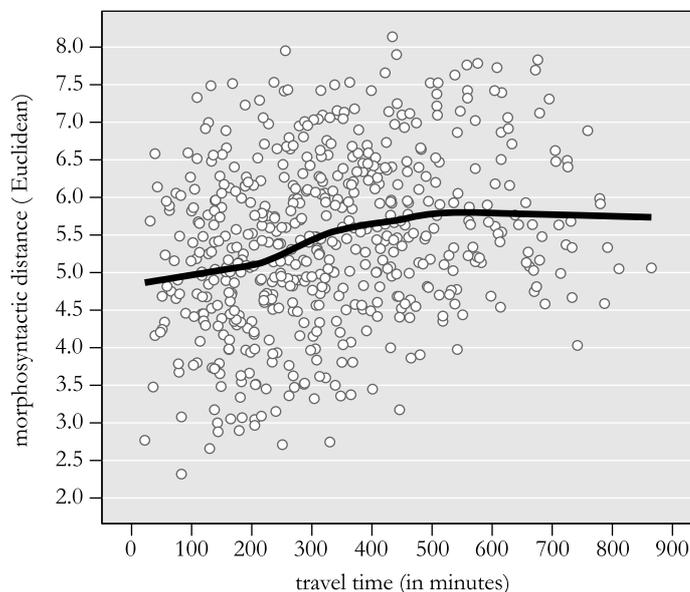
Figure 2: Scatterplot – morphosyntactic distances (vertical axis) vs. least-cost travel time (horizontal axis). Solid line: non-parametric regression curve. Explained variance (logarithmic estimate): $R^2 = .076$ ($p < .001$)

from speakers born around the beginning of the 20th century with travel estimates based on 21st century transportation infrastructure is obviously an issue – but given that transportation infrastructure grows rather organically, we do not expect this temporal mismatch to distort results too severely.

Correlating the least-cost travel time matrix with the original Euclidean morphosyntactic distance matrix yields the scatterplot in figure 2. Once again, we find a positive relationship such that increased travel time typically entails increased morphosyntactic distance. Having said that, the regression curve in the diagram suggests that this time, we are indeed dealing with a sublinear relationship (cf. Séguy 1971): morphosyntactic distance increases with travel time until a travel time of approximately 500 minutes (a little more than eight hours) – after that, the curve levels off. In other words, after eight hours of traveling, dialects to be encountered at later stages of the journey will not be substantially more distant from the dialect at the origin of the journey. Statistical analysis confirms that a logarithmic estimate of the relationship fits the data best, explaining ca. 7.6 % of the variance in morphosyntactic distances ($R2 = .076$, $p < .001$).

In sum, least-cost travel time is a better predictor of morphosyntactic distances than as-the-crow-flies distance, explaining 3.2 per cent points more of

the linguistic variability. Having said that, the predictors' explanatory power is still rather limited.

4.3. Linguistic gravity

In a (1974) paper, Peter Trudgill suggested a gravity model to account for geographic diffusion. Trudgill argued 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). As a testable hypothesis, this formula[4] – which, as its name says, is inspired by Newton's famous inverse-square law of gravitation – postulates that the interaction between two dialects decreases with increasing geographic distance (in our parlance, that the linguistic distance between two dialects increases with increasing geographic distance), but that this effect is counterbalanced by larger speaker communities: large speaker communities will tend to linguistically interact more than smaller speaker communities, all other things (and especially geographic distance) being equal. To fix terminology, the numerical value for the interaction effect suggested by Trudgill will henceforth be referred to as *Trudgill's linguistic gravity index* (or TLGI for short). Observe, in this connection, that previous dialectometrical research failed to detect linguistic gravity effects in dialect data. Will we find a gravity effect in our data?

The scatterplot in figure 3 suggests that we do. Using Trudgill's formula, we calculated TLGI values for every one of the 561 county/county pairings in our database, feeding in least-cost travel time as geographic distance measure and early 19th century population figures[5] (in thousand) as a proxy for speaker community size. The population figures thus roughly match the date of birth of most FRED speakers. Figure 3, then, confirms the theoretically expected relationship between TLGI values and morphosyntactic distances: increased gravity (in Trudgill's diction, increased *interaction*) between two dialects predicts smaller morphosyntactic distance. The best estimate of

----

[4] Subsequently in the paper, Trudgill amends this equation in several ways. For one thing, he introduces a linguistic similarity variable *s* into the equation. This variable, however, will not be considered in the present study, as it is precisely linguistic (dis)similarities that we seek to explain by the gravity concept. Later in the paper, Trudgill also alters the formula to measure *influence*, not *interaction* – again, we shall ignore this amendment because here we are interested in interaction, not influence.

[5] Specifically, we used 1901 figures, as published in the *Census of England and Wales, 1921* and the *Census of Scotland, 1921*. These documents are available online at http://histpop.org/.
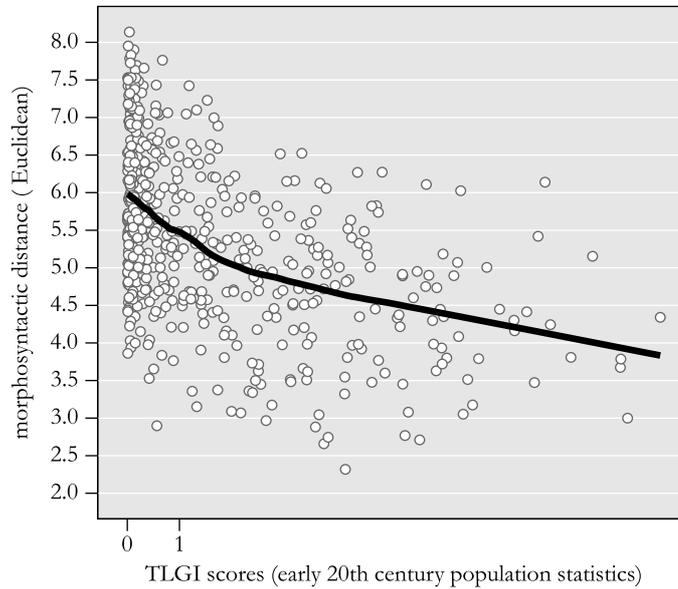
Figure 3: Scatterplot – morphosyntactic distances (vertical axis) vs. Trudgill's linguistic gravity indices (TLGI) scores (horizontal axis, log scale). Solid line: non-parametric regression curve. Explained variance (logarithmic estimate): $R^2 = .241$ ($p < .001$)

this relationship is logarithmic (i.e. sublinear) and yields an $R2$ value of .241 ($p < .001$). In other words, linguistic gravity explains about 24% of the observable variance in morphosyntactic distances.

We should add a word on the logarithmic nature of the relationship. A more detailed analysis reveals that the relationship is fairly linear for the 464 data points in the sample with TLGI values smaller than 10 (the mean TLGI value in the dataset is 16.71). However, TLGI values larger than 10 (this concerns the remaining 97 data points) do not coincide with proportionally smaller morphosyntactic distances. Consider the pairing between the counties Middlesex and London: given its huge population product (792,000 × 4,536,000) and its comparatively small squared travel time (55 min × 55 min), this pairing has a truly enormous TLGI value (1188), but needless to say this does not mean that morphosyntactic distance between the two counties is just a hundredth of the distance we observe for the 464 pairings with TLGI values that are smaller than 10 (as a matter of fact, the Middlesex-London pairing has a morphosyntactic distance of 4.34; the data points with TLGI values < 10 are associated with a mean morphosyntactic distance of 5.60). This is why the regression curve in Figure 3 levels off towards the right half of the diagram.

We conclude that the notion of linguistic gravity, which blends geographic information with population statistics, has a significant effect on morphosyntactic distances. This effect is responsible for a quarter of the variance in morphosyntactic distances.

4.4. Dialect groupings and dialect areas

Given that we have detailed earlier how mere geographic distance appears to be somewhat irrelevant for predicting morphosyntactic distance, this section is concerned with testing the dialect area view (cf. Heeringa & Nerbonne 2001) for its explanatory potential. To partition our data points into discrete groups that are internally coherent linguistically, we utilize customary hierarchical agglomerative cluster analysis (cf. Aldenderfer & Blashfield 1984).[6]

Cluster analysis yields tree diagrams, also known as dendrograms, where one finds individual varieties to the left and successively larger clusters as one moves rightwards. Essentially, dendrograms work in much the same way as family trees. A consensus dendrogram derived from the present study's dataset is displayed in figure 4. We find three fairly robust clusters: the first one – Angus (ANS) through Ross and Cromarty (ROC) – comprises all the Scottish dialects plus some enclaves in England (Northumberland, Warwickshire, Middlesex) and Wales (Denbighshire). The second cluster – Cornwall (CON) through Suffolk (SFK) – groups Southern English English dialects with Durham (DUR) in the North of England. The third cluster – Glamorganshire (GLA) through Shropshire (SAL) – unites Northern English English dialects and Glamorganshire in Wales. All said, then, we diagnose some geographic incoherence in the dialect groupings identified here.

How much of the observable variance in morphosyntactic distances can be accounted for by taking into account these dialect groupings? To provide an answer to this question (we hasten to add that there are other, statistically more sophisticated answers), we will now correlate morphosyntactic distances with cophenetic distances, that is, with consensus distances that are

------

[6] On a technical note, we used Ward's Minimum Variance Method, a popular algorithm – also in dialectometry – that tends to create small and even-sized clusters. Because simple clustering can be unstable, a procedure known as 'clustering with noise' (Nerbonne, Kleiweg & Manni 2008) was applied: the original Euclidean distance matrix was clustered repeatedly (in our case, 10,000 times), adding a random amount of noise in each run. This yields a cophenetic distance matrix which details consensus (and thus more stable) cophenetic distances between localities, and which is amenable to various visualization and analysis techniques. We used the noise ceiling ($c = s/2$) suggested in Nerbonne et al. (2008).
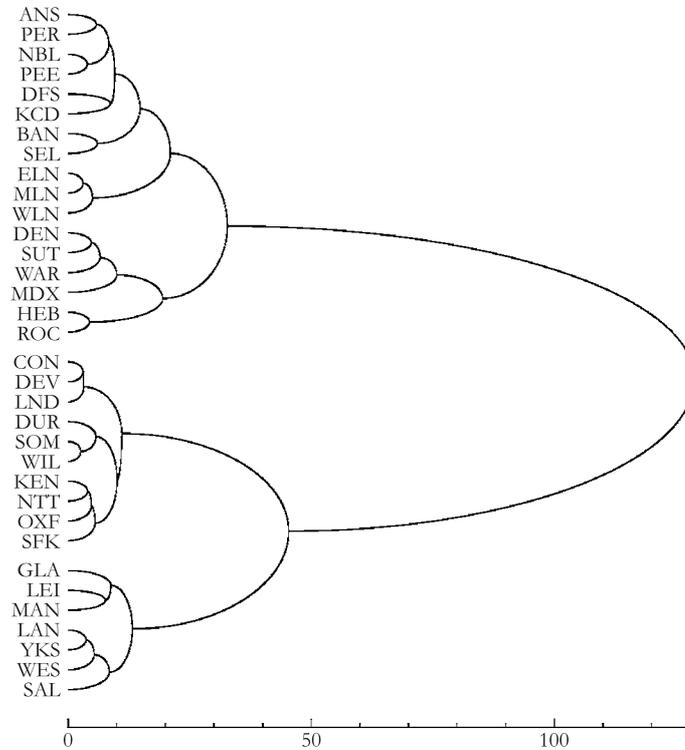
Figure 4: Consensus dendrogram – hierarchical agglomerative cluster analysis, clustering algorithm: Ward's method

the basis for the dendrogram in figure 4 (cf. Nerbonne 2008). Succinctly put, a cophenetic distance is the distance you have to travel in a dendrogram to get from one node to another node. The idea is that while your brother may live on the other side of the world and your third grade cousin next door, in terms of your blood line (or family tree) you are still a lot closer to your brother than your cousin, geographic distance notwithstanding. In the same vein, two dialects may be distant geographically but still very similar linguistically, as is the case with, e.g., Middlesex and Sutherland. In this spirit, the scatterplot in figure 5 plots morphosyntactic distances against cophenetic distances. The plot makes amply clear that, first, there is a quite robust positive relationship such that increased cophenetic distance implicates increased morphosyntactic distance. Second, notice that the regression curve levels off towards the right half of the dendrogram. So once again, we are dealing with a sublinear relationship such that after a certain threshold, larger cophenetic distances do not lead to proportionally larger morphosyntactic
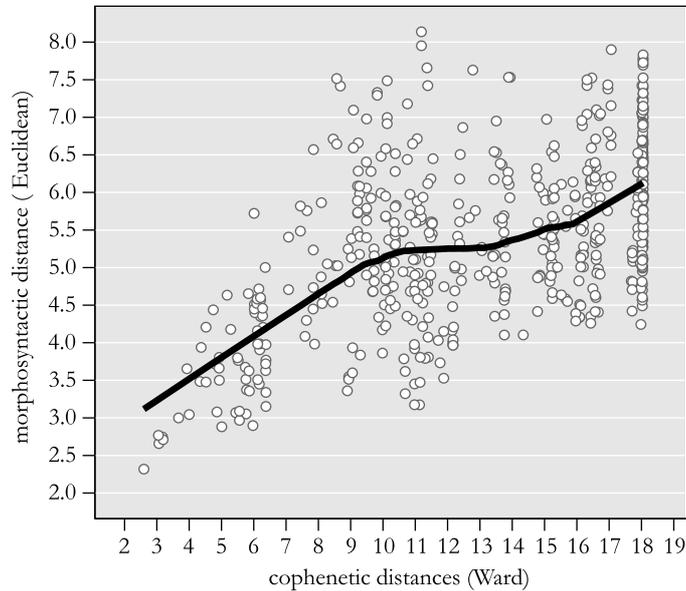
Figure 5: Scatterplot – morphosyntactic distances (vertical axis) vs. cophenetic distances (horizontal axis; clustering algorithm: Ward's method). Solid line: non-parametric regression curve. Explained variance (logarithmic estimate): $R^2 = .325$ ($p < .001$)

distances. It is therefore not surprising that a logarithmic estimate of the relationship fares slightly better ($R2 = .325$, $p < .001$) than a linear one ($R2 = .297$, $p < .001$).

We have seen in this section that dialect groupings appear to be the most powerful explanatory variable considered in this paper: taking them into account enables us to account for roughly a third of the observable variance in morphosyntactic distances.

5.   Summary and conclusion

The aim of this paper was to probe the validity of the FDP, a time-honored principle according to which geographic proximity predicts dialectal similarity. It is fair to say that the FDP has failed this test. Utilizing an aggregational, frequency-based approach to morphosyntactic variability in British English dialects, we have seen that as-the-crow-flies distance accounts for just 4 % of the morphosyntactic variance, and least-cost travel time for about 8 %. Blending geography with population statistics yields more explanatory power: thus, Trudgill's notion of linguistic gravity can explain circa 24 % of

the variability in the data. Finally, partitioning dialects into (geographically not necessarily coherent) groupings – the dialect area view – and correlating those dialect groupings with morphosyntactic distances accounts for roughly 33 % of the morphosyntactic variance. Plainly, therefore, geographic proximity or distance *per se* does not explain a whole lot. In other words, the FDP does not mesh well with the facts and overrates the role that geography play.

With previous estimates for the explanatory potency of geography typically exceeding 30 % and sometimes approaching 80 %, it is fair to ask how it is possible that this paper finds the FDP to lack explanatory value. Three scenarios are theoretically possible:

– *Morphosyntactic variability in British English dialects is different from morphosyntactic variability elsewhere.* Unlikely – there is some (albeit not much) previous dialectometrical research on English dialects drawing on morphosyntactic features (e.g. Goebl 2007), and this research has detected clear geographic proximity signals.

– *Morphosyntactic variability – and especially syntactic variability – is different from accent or lexical variability.* This explanation is more convincing, given a long-held suspicion by typologists, syntacticians and even dialectometricians that morphosyntax is less amenable to geographic diffusion than e.g. pronunciational variability. Still, recall that Spruit, Heeringa & Nerbonne (2009), in their study on syntactic differences between Dutch dialects, find that as-the-crow-flies distance explains a lot (45 %) of observable syntactic variability.

– *Compared to frequency-based approaches, atlas-based approaches overrate geography.* For two reasons, this is the most cogent scenario. For one thing, feature selection does matter a great deal, and it is fair to ask 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. As an aside, we note in this connection that if we had based our analysis in this paper on only those 19 features in our catalogue that have a significant geographic distribution, we would have been able to account for 19 % of the linguistic variance by considering as-the-crow-flies distance alone. Second, we cannot rule out at this point that the categorical (attested vs. not attested) and mediated (by fieldworkers) second-hand nature of the information available in linguistic atlases also systematically overrates geography, in that contrasts and distinctions appear as more pronounced than they actually are.

The last scenario, in particular, raises a number of empirically and theoretically interesting questions – what is the most appropriate method to assess how things 'actually are'? – but a satisfactory discussion of the many issues involved here is, alas, beyond the scope of the present paper. The jury is still out on the relationship between frequency-based and atlas-based dialectometry.

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-´* (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 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 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*)
  [43] 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

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

Anderwald, Lieselotte and Benedikt Szmrecsanyi   2009 Corpus linguistics and dialectology. In: Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics. An International Handbook*, 1126–1139. Berlin/New York: Mouton de Gruyter.

Cronbach, Lee J.   1951 Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.

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

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

Goebl, Hans   1993 Probleme und Methoden der Dialektometrie: Geolinguistik in globaler Perspektive. In: Wolfgang Viereck (ed.), *Proceedings of the International Congress of Dialectologists*, 37–81. Stuttgart: Steiner.

Goebl, Hans   2007 A bunch of dialectometric flowers: a brief introduction to dialectometry. In: Ute Smit, Stefan Dollinger, Julia Hüuttner, Gunter Kaltenböck and Ursula Lutzky (eds.), *Tracing English through time: Explorations in language variation*, 133–172. Wien: Braumüller.

Heeringa, Wilbert and John Nerbonne   2001 Dialect areas and dialect continua. *Language Variation and Change* 13: 375–400.

Hernández, Nuria   2006 *User's Guide to FRED.* Available online at http://www.freidok.uni-freiburg.de/volltexte/2489/. Freiburg: English Dialects Research Group.

Kortmann, Bernd and Benedikt Szmrecsanyi   2004 Global synopsis: morphological and syntactic variation in English. In: Bernd Kortmann, Eddgar Schneider, Kate Burridge, Rajend Mesthrie and Clive Upton (eds.), *A Handbook of Varieties of English*, Volume 2, 1142–1202. Berlin/New York: Mouton de Gruyter.

Nerbonne, John   2006 Identifying Linguistic Structure in Aggregate Comparison. *Literary and Linguistic Computing* 21: 463–475.

Nerbonne, John   2008 Areal effects in varietal classification. Talk given at *NWAV37* Houston, TX, November 6–9, 2008.

Nerbonne, John   (to appear) Various variation aggregates in the LAMSAS South. In: Catherine Davis and Michael Picone (eds.), *Language Variety in the South III.* Tuscaloosa: University of Alabama Press.

Nerbonne, John, Wilbert Heeringa and Peter Kleiweg   1999 Edit Distance and Dialect Proximity. In: David Sankoff and Joseph Kruskal (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison.* Stanford: CSLI Press, x-vx.

Nerbonne, John and Peter Kleiweg   2007 Toward a Dialectological Yardstick. *Journal of Quantitative Linguistics* 14: 148–166.

Nerbonne, John, Peter Kleiweg and Franz Manni   2008 Projecting dialect differences to geography: bootstrapping clustering vs. clustering with noise. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt and Reinhold Decker (eds.), *Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society*, 647–654. Berlin: Springer.

Orton, Harold and Eugen Dieth   1962 *Survey of English Dialects.* Leeds: E. J. Arnold.

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

Shackleton, Robert G. Jr.   2007 Phonetic variation in the traditional English dialects: a computational analysis. *Journal of English Linguistics* 35: 30–102.

Spruit, Marco René, Wilbert Heeringa and John Nerbonne   2009 Associations among Linguistic Levels. *Lingua* 119(11): 1624–1642.

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

Szmrecsanyi, Benedikt   2010 *The morphosyntax of BrE dialects in a corpus-based dialectometrical perspective: feature extraction, coding protocols, projections to geography, summary statistics.* URN: urn:nbn:de:bsz:25-opus-73209, URL: http://www.freidok.uni-freiburg.de/volltexte/7320/. Freiburg: University of Freiburg.

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

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

Wälchli, Bernhard   2009 Data reduction typology and the bimodal distribution bias. *Linguistic Typology* 13: 77–94.

Benedikt Szmrecsanyi
Geography is overrated
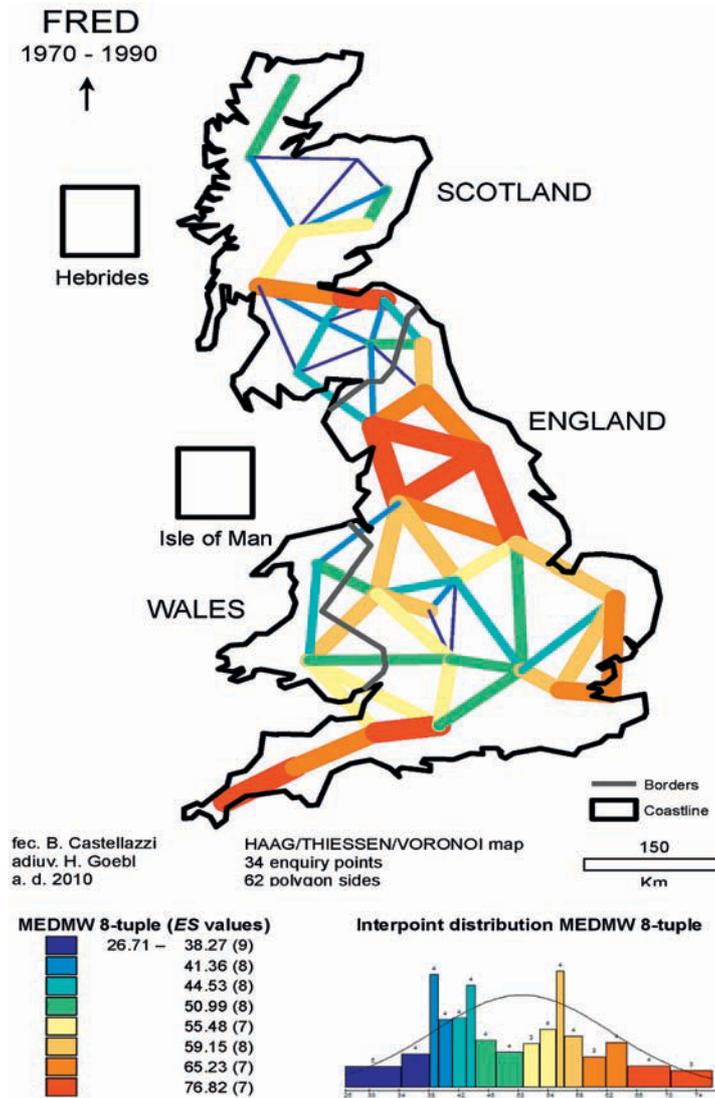
Map 1: Beam map. Neighbors that are close morphosyntactically are connected by warm and heavy beams; morphosyntactically dissimilar neighbors are linked by cold (blueish) and thin beams.