Chapter 13

Top-down and bottom-up advances in corpus-based dialectometry

Christoph Wolk
University of Giessen

Benedikt Szmrecsanyi
KU Leuven

We present three approaches to corpus-based dialectometry and apply them to morphosyntactic variation in the *Freiburg Corpus of English Dialects*, which covers 34 counties throughout Great Britain. Two of these are *top-down approaches* that start with a predefined feature list; one using a straightforward frequency-based analysis, the other enhancing the raw numbers using probabilistic modeling. Both methods are able to detect the structure of areal variation in Great Britain, and the second approach is able to reduce the influence of textual coverage as a nuisance factor. The final approach is a *bottom-up* method that eschews pre-specified lists and evaluates potential features directly from the data using a permutation-based metric. Again, we find that simple frequency-based metrics are biased, but that derived metrics yield a clearer pattern. Using these methods, we are able to uncover significant geolinguistic structure in Great Britain.

1 Introduction

In this contribution, we sketch novel ways to conduct dialectometry. Let us set the scene by fixing some terminology first. LINGUISTIC CORPORA are principled and broadly representative collections of naturalistic texts or speech. Linguistic corpora thus sample USAGE DATA, and as such are increasingly popular in dialectology (Anderwald & Szmrecsanyi 2009; Grieve 2009) and beyond (see the papers in Szmrecsanyi & Wälchli 2014). CORPUS LINGUISTICS, accordingly, is a methodology in linguistics that bases claims about language on corpora. Corpus linguistics is thus the methodological outgrowth of the usage-based turn in linguistics, in the spirit of, e.g., Bybee (2010); Tomasello (2003). As is well known, CLASSICAL

Christoph Wolk & Benedikt Szmrecsanyi. 2016. Top-down and bottom-up advances in corpus-based dialectometry. In Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.), *The future of dialects*, 225–244. Berlin: Language Science Press. DOI:10.17169/langsci.b81.152

DIALECTOMETRY in the tradition of Goebl (1984); Nerbonne, Heeringa & Kleiweg (1999) draws on atlas material to explore geolinguistic patterns using aggregation methodologies. By contrast, CORPUS-BASED DIALECTOMETRY (henceforth: CBDM) utilizes aggregation methodologies to explore quantitative and distributional usage patterns extracted from dialect corpora.

Why do we need CBDM? After all, there is some scepticism in the community about the usefulness of non-atlas resources (Goebl 2005a: 499, for example, writes that "Extra atlantes linguisticos nulla salus dialectometrica"). Let us emphasize first that we do not wish to suggest that linguistic atlases are dispensable. Quite on the contrary, we are convinced that they are quite indispensable for some purposes, such as surveying – from a bird's eye perspective – the variable presence or absence of particular features in particular language or dialect areas. But at the same time we believe that also (!) being able to analyze naturalistic corpus data is central to the maturation of the dialectometry enterprise. The reason is that the data in (most) linguistic atlases speak primarily to the issue of explicit, active linguistic knowledge, while corpus data document first and foremost usage (which is, of course, related to knowledge of the more implicit sort, but there is no 1-1 correspondence). Thus turning to corpora will enable dialectologists to address hitherto rather neglected questions about usage versus knowledge, production/comprehension versus intuition, chaos versus orderliness, and so on. We should also add that neighboring linguistic disciplines, such as variationist sociolinguistics, rely empirically almost exclusively on usage data. It is precisely because of this that there is some welcome methodological convergence to be had in the field. The disadvantage of CBDM is that corpus methodologies are not well suited to study low-frequency phenomena; rare things, in other words, are better investigated drawing on atlases and surveys.

In this contribution we sketch three CBDM approaches, two top-down, the third bottom-up. The top-down approaches first define a feature catalogue, then establish the frequencies and/or probabilities associated with the features, and finally aggregate over them. The bottom-up approach, by contrast, lets the features to be aggregated emerge in a data-driven fashion through the identification of significant and/or distinctive part-of-speech *n*-grams. The case studies which we present to illustrate these approaches summarize work by Szmrecsanyi (2013) and Wolk (2014). All case studies are concerned with grammatical variation in traditional British English dialects.

This contribution is structured as follows. In §2 we sketch the dialect corpus into which we tap. §3 describes the top-down CBDM approach; §4 is dedicated to the bottom-up approach. §5 offers some concluding remarks.

2 The Freiburg Corpus of English Dialects (FRED)

The case studies in this contribution will analyze the *Freiburg Corpus of English Dialects* (henceforth: FRED) (see Hernández 2006 for details). The version of the corpus used in the top-down CBDM study (§3) contains 368 individual texts and covers approximately 2.44 million words of running text (this corresponds to about 300 hours of speech), mainly transcribed so-called 'oral history' material. These were mostly recorded between 1970 and 1990. The typical setting is that a fieldworker interviews an informant about life, work, etc. in the olden days. The 431 informants represented in the corpus are typically elderly people with a working-class background – so-called NORMS (*non-mobile old rural males*) (cf. Chambers & 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. These counties are displayed in Figure 1. Note that we aggregate all texts/interviews from the same county, in order to obtain 34 distinct subcorpora. Individual texts are annotated with longitude/latitude information; county coordinates (mean longitude and latitude) have consequently been calculated by computing the arithmetic mean of all the location coordinates associated with a particular county. Some of the informants had to be removed for the analysis presented in §3.2 due to missing metadata. This led to the complete removal of three counties for that analysis: East Lothian, Denbighshire and Warwickshire.

In the bottom-up CBDM study (§4), we will analyze a smaller version of the corpus: the Freiburg Corpus of English Dialects Sampler (FRED-S) (Szmrecsanyi & Hernández 2007). FRED-S contains a subset of the texts in the full FRED corpus totaling about 1 million words of running text and covering 17 counties in England and the Scottish Lowlands. The big advantage of using FRED-S, though smaller in size, is that it exists in a version that was automatically part-of-speech (POS) annotated by the CLAWS4 tagger (Garside & Smith 1997) using the detailed CLAWS7 tagset.

To illustrate the nature of the material sampled in FRED and FRED-S, (1) is the beginning of an interview conducted in 1978 in St. Ives, Cornwall (FRED text CON003). The informant is an 86 year-old male ('CAVA_PV'), who is interviewed by two interviewers ('IntRS' and 'Inf'). Interviewer utterances are enclosed in curly brackets (note that interviewer utterances are excluded from analysis in the present study).

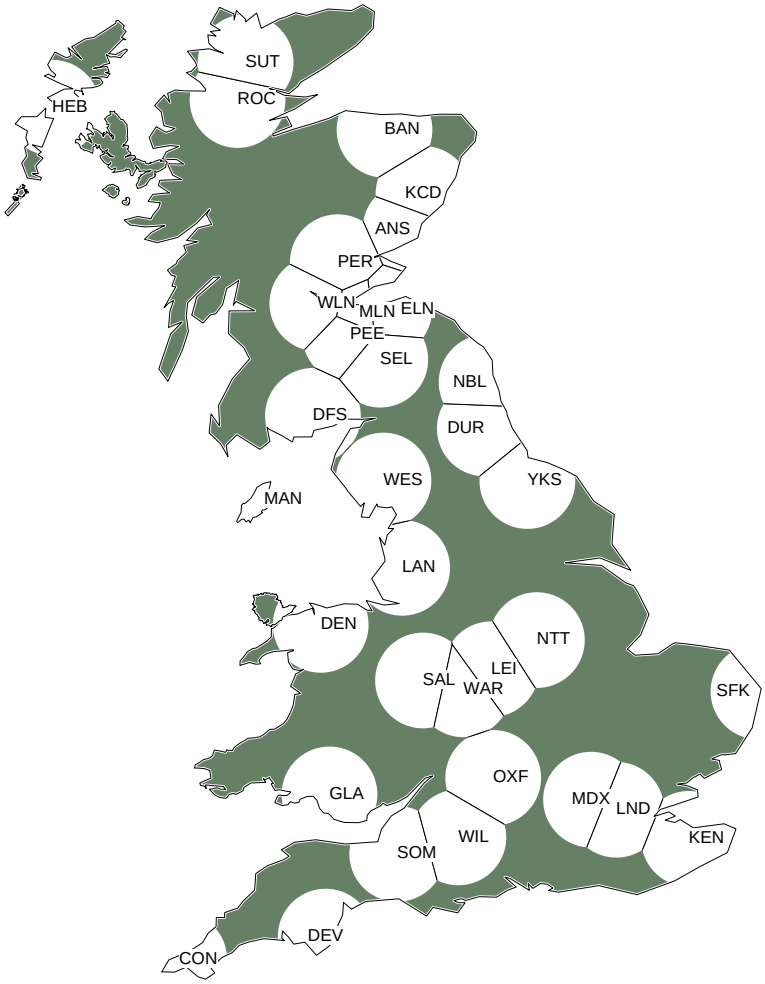

Figure 1: Pre-1974 counties represented in FRED. See http://en.wikipedia.org/wiki/Chapman_code for an explanation of the county codes.

(1) *{<u IntRS> Well you're a St. Ives man. Where were you born?}*
<u CAVA_PV> Born Belyars Lane, eighteen ninety-two. Eighteenth of December. Worn sovereign in the cupper. Born sovereign. The poor times then, you know (gap 'indistinct') boiling potatoes and t – inkle mosses.
{<u IntRS> Did you, did you, how long did you live there?}
<u CAVA_PV> Oh we lived there about, oh about twelve years, I suppose. Then we went up to a Rosewall Terrace. Hmm. So everything's altered now to what er was then, I mean.

3 Top-down CBDM

This section will discuss top-down CBDM, which consists of five steps:

Step 1: define the feature catalogue (motto: the more features, the merrier).

Step 2: identify features in the corpus texts (automatically, semi-automatically, or manually).

Step 3: establish raw feature frequencies (per location); subsequently, normalize frequencies and/or model frequencies probabilistically.

Step 4: aggregate: calculate a distance matrix.

Step 5: project to geography, analyze & interpret.

Szmrecsanyi (2013) and Wolk (2014) discuss the method in meticulous detail. Suffice it to say here that the feature catalogue we used to explore grammatical variation in British English dialects consists of $p = 57$ features, which cover all major domains in English grammar as well as the usual suspects in the variationist and dialectological literature, such as non-standard past tense *done* (e.g., *you came home and done the home fishing*), multiple negation (e.g., *don't you make no damn mistake*), and *don't* with third person singular subjects (e.g., *if this man don't come up to it*). These features were identified in the corpus material (automatically, semi-automatically, or manually – depending on the nature of the feature), and their usage frequency established. Subsequently, this information was arranged in an n by p table: 34 counties, each characterized by a vector of 57 feature frequencies. At this point, there are two ways to proceed: the *normalization-based top-down CBDM approach*, pursued in Szmrecsanyi (2013), and the *probabilistically enhanced top-down CBDM approach*, explored in Wolk (2014). We will now discuss these top-down variants in turn.

3.1 The normalization-based top-down CBDM approach

Szmrecsanyi (2013) processed the frequency table in two ways prior to analysis. For one thing, he normalized raw frequencies to frequency per 10,000 words, doing justice to the fact that textual coverage of individual dialects varies. This normalized frequency table tells us, for example, that multiple negation is twice as frequent in Nottinghamshire than in Yorkshire. Additionally, Szmrecsanyi applied a log-transformation to the normalized frequencies for the sake of deemphasizing large frequency differentials and thus alleviating the effect of frequency outliers. Next Szmrecsanyi converted the normalized, log-transformed frequency table into an N by N distance matrix. This transformation is an aggregation step, in that the resulting distance matrix abstracts away from individual feature frequencies and specifies pairwise distances between the objects considered. To calculate dialectal distances, Szmrecsanyi used the well-known Euclidean Distance Measure (see, e.g., Aldenderfer & Blashfield 1984: 25). The Euclidean Distance Measure defines the distance between two dialects as the square root of the sum of all p squared frequency differentials.

Distance matrices are the customary input to dialectometric analysis and visualization techniques. Let us now explore the normalization-based top-down distance matrix using a particularly popular dialectometric mapping technique, *cluster maps*. Cluster maps are common in all strands of dialectometry, and project the outcome of cluster analysis to geography (cf., for example, Goebl 2007: Map 18; Nerbonne & Siedle 2005: Figure 5; Heeringa 2004: Figure 9.6). First, an N by N distance matrix is subjected to HIERARCHICAL AGGLOMERATIVE CLUSTER ANALYSIS (cf. Jain, Murty & Flynn 1999; we specifically used Ward's method as the clustering algorithm), a statistical technique used to group a number of objects (in this study, dialects) into a smaller number of discrete clusters.[1] Cluster memberships of dialect locations can then be projected to geography via, e.g., color coding.

In the left-hand and center maps of Figure 2 we find two cluster maps that correlate (Goebl 2005b) Great Britain's geographic landscape to its dialect landscape. For expository purposes, both maps display a 3-cluster solution, but we do not wish to claim that this is necessarily the optimal solution. The left-hand map clusters a distance matrix detailing not linguistic distances but as-the-crow-flies geographic distances between dialect sites, thus depicting, for reference purposes, a geographically maximally neatly partitioned map. This map suggests that on strictly geographic grounds, Great Britain can be partitioned into three coherent

[1] Simple clustering can be unstable, so we used the "clustering with noise" technique (Nerbonne et al. 2008).

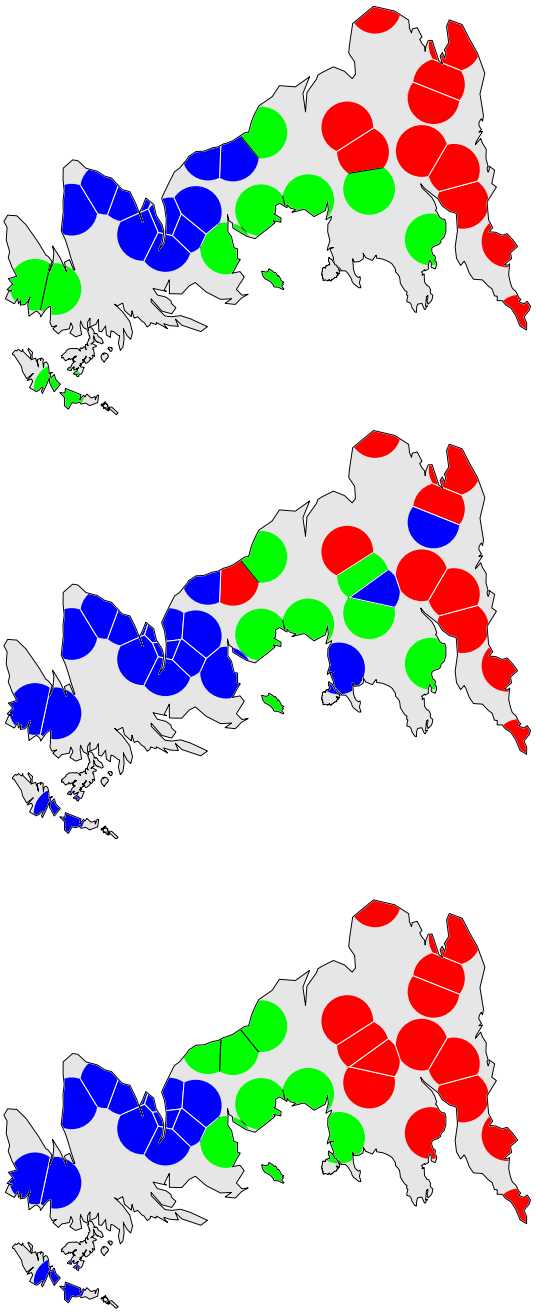

Figure 2: Cluster maps – hierarchical agglomerative cluster analysis (displayed: 3-cluster solution). Left: Geographic as-the-crow-flies distances. Center: Normalization-based morphosyntactic distances. See Szmrecsanyi (2013: chapter 6) for details. Right: Probabilistically enhanced morphosyntactic distances.

areas: a red region comprising the South of England plus the county of Glamorganshire in Southern Wales; a green region containing the North of England plus the county of Denbighshire in Northern Wales plus the county of Dumfriesshire in Southern Scotland; and a blue region encompassing Scotland minus the county of Dumfriesshire.

Compare this scenario to the map in the center of Figure 2, which projects a corresponding regionalization on morphosyntactic grounds. There is a good deal of geographic incoherence in the morphosyntax division: For example, there are blue outliers all over England and Wales; Durham in the North of England is categorized as a red (i.e. Southern) county; Glamorganshire in Southern Wales is a green (i.e. Northern) county; and so on. That said, there is clearly some similarity between the geographic and linguistic partitioning, because the tripartite division between Scotland, the North of England, and the South of England is essentially in place. We conclude that our corpus-based measure of aggregate morphosyntactic variability does detect a geolinguistic signal.

3.2 The probabilistically enhanced top-down CBDM approach

While the signal discussed in the previous section seems to broadly match the description in the literature, it also raises some concerns. First, the outliers are difficult to motivate. Why should, for example, Middlesex group with Scotland? For most of the outliers, individual significant differences to their geographically close neighbors can be found (Szmrecsanyi 2013: chapter 7), but this does not sufficiently explain the cluster structure. Second, the results do not confirm two of the most reliable results of the atlas-based dialectometric enterprise: both the shape and the strength of the relationship between linguistic and geographic distances are markedly different. Nerbonne (2013) summarizes several studies, finding that geographic distance (statistically) explains 16 to 37 percent of the variance in linguistic distance, and that the relationship is sublinear: as one considers location pairs that are further apart, the increase in linguistic dissimilarity begins to level off. In contrast the corpus signal yields a very low correlation between linguistic and geographic ("as the crow flies") distances, explaining only approximately 4.4 percent of the variance, and the relationship is linear rather than sub-linear. Using travel time as the operationalization of geographic distance (Gooskens 2005) improves the relationship slightly to almost 8 percent; nevertheless, it is still far below what is typically found. This suggests that some form of bias may exist in the data set.

It is a well-known effect that non-linguistic aspects of the data set and its creation can influence the aggregate results, such as the specific fieldworker cover-

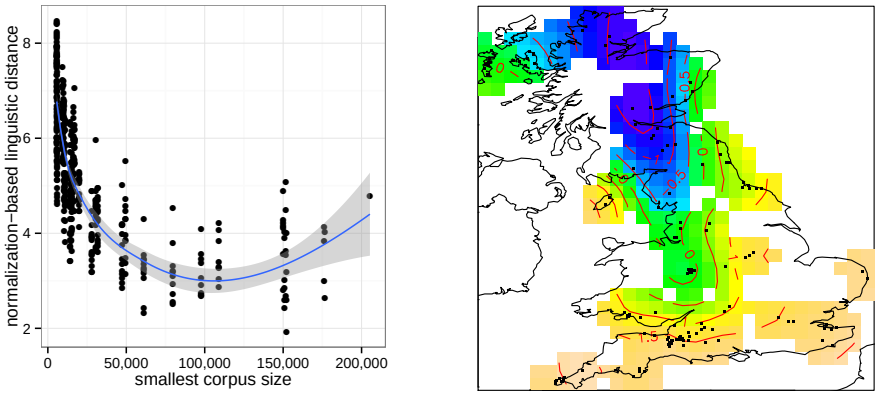

Figure 3: Left: Correlation of minimum subcorpus size and linguistic distance (top down normalization-based approach; $R^2 = 0.61$) Right: Example GAM for multiple negation (log scale). Lighter colors indicate higher frequency. Red lines indicate shape of the frequency gradient.

ing a location ("field worker isoglosses", Trudgill 1982: 241ff.). Similar problems may reside in the corpus at hand. We suggest that one issue in particular causes a substantial amount of the divergences from the usual pattern: the fact that the amount of corpus material per county varies, and in some cases the number of words per county may be very small. Measurements that are based on little data are imprecise, and so are the distances resulting from them. We can test the influence of this factor by exploring the relationship between linguistic distance and an appropriate operationalization of corpus size (and therefore accuracy).

The left-hand part of Figure 3 displays the linguistic distance between county pairings as a function of the smaller of the two subcorpus sizes; a smoother line is included to highlight the general trend. Clearly, there is a strong relationship: distances involving the counties with the worst coverage are consistently too high. At approximately 50,000 words, this relationship largely levels off.[2] The outliers in Figure 2 fall below this threshold. For example, the subcorpus for Durham consists of 28,000 words, and that for Middlesex of 32,000. These three measures of quality – interpretability of groupings, correlation with geography, and influence of the amount of data – suggest that there may be a bias in the data that obscures the pattern. Note, however, that these measures serve more as

[2] The small uptick at the end is a combination of data sparsity (due to the small number of subcorpora of that size) and the somewhat atypical, but large Suffolk subcorpus.

"sanity checks"[3] than as proper external validation. A method that fares better under this yardstick is not necessarily correct, but a method that fares worse indicates potential problems.

We therefore propose to use some form of smoothing that takes the accuracy of the measurement into account. Per the Fundamental Dialectological Postulate (Nerbonne & Kleiweg 2007), geographic smoothing seems particularly appropriate: in the absence of compelling evidence to the contrary, we should assume that proximate varieties resemble each other. Several methods of doing such smoothing have been proposed, including intensity estimation (Rumpf et al. 2009), local spatial autocorrelation (Grieve 2009), and generalized additive models (GAMs; Wieling 2012). We believe that GAMs are an especially adequate choice, as they are a variant of regression modeling, and therefore closely resemble the techniques in common use by, among others, variationist sociolinguists (Tagliamonte 2012). Using such models, it is possible to account for other factors that may influence the result, whether language-internal (e.g. subject type) or -external (e.g. speaker age). In Wolk (2014), two language-external factors were included, namely speaker age and gender.[4] We keep these two factors for the analysis presented here, to account for any imbalances in the corpus sampling process. For many features, there are significant effects of these factors, largely in the expected direction (i.e. female speakers use fewer non-standard features and older speakers use more archaic features). Simulations based on these results, however, suggest that the overall effect that the inclusion of these factors has on the resulting distances is marginal (Wolk 2014: 233f.).

In contrast to the GAM-based method used in Wieling (2012), Wolk (2014) did not model the distances directly, but built a separate model for each feature. Furthermore, features that represent binary alternations,[5] such as habitual *would* vs. *used to* or *will* vs. *going to* as future markers, were modeled as such, rather than as individual frequencies. Doing so removes potential bias resulting from base rate differences (e.g. differing frequencies of habitual or future contexts regardless of form) between speakers/counties, and makes the results more comparable to variationist research on these features, which typically utilizes VARBRUL-style modeling. This yields a list of 45 remaining features, and therefore the number of models included in the analysis is also 45. Each model contains a two-dimensional geographic smoother that allows the feature to vary by location in

[3] We thank an anonymous reviewer for this phrasing.

[4] As the relevant information was missing for a small number of texts (including all texts from three counties, see §2), they had to be removed from the analysis.

[5] The selection of features to model as alternations was based on the variationist literature and on certain features existing as standard/non-standard variants in the original list.

a gradient fashion. An example for this can be seen in the right-hand map in Figure 3, displaying the smoother for multiple negation. As expected, the feature is rare in Scotland and relatively frequent in the South of England, with the North forming a transition zone. This model is then used to predict the proportion of one realization, or the frequency to be expected in ten thousand words. From here on, the analysis proceeds as outlined above.

A cluster map of the resulting distances is presented in the right-hand part of Figure 2. The tripartite division remains, and the large-scale areas are geographically coherent, with the exception of the young dialects in the Scottish Highlands and the Hebrides, which group with the North of England. Quantitatively, the model fares well. The relationship between geographic and linguistic distances is sublinear and solid ($R^2 = 0.44$ for least-cost travel time; compare the normalization-based value of 0.08). Even more importantly, the influence of subcorpus size has greatly decreased and now accounts for only 16.2 percent of the variance, compared to 61 percent for the normalization-based distances.

We hasten to add that this clean pattern is hardly surprising: the GAMs have geographic coherence built into their assumptions. Therefore, it can be argued that the results may be too homogeneous, that there are true differences that the GAMs smooth over. Nevertheless, the model produces at least an upper boundary for spatial cohesion; and a bias in favor of the Fundamental Dialectological Postulate seems more plausible than a bias toward accidental properties of the corpus compilation process. Furthermore, the resulting association between geography and linguistic distance, while on the high side, is not outside the range of what one would expect based on traditional dialectometric analyses. It is also clearly distinct from 1, indicating that there is still unpredictable dialectal variation left. Finally, it bears noting that the GAM process yields interpretable single feature maps as a byproduct – a rather beneficial property of this approach.

4 Bottom-up CBDM

So far, we have covered methods that rely on a pre-specified feature list. In the following, we explore whether it is possible to eliminate such a list and go directly from the corpus to a distance measurement. Directly measuring corpus (dis)similarity is an important, yet somewhat underresearched, topic in corpus linguistics (Kilgarriff 2001). This is especially true for morphosyntactic measures.[6] The method proposed here builds on an idea proposed by Nerbonne &

[6] Scherrer (2012) provides a method for deriving pronunciation distances between corpora automatically, but this approach is not straightforwardly generalizable to morphosyntax.

Wiersma (2006), who employ part-of-speech *n*-grams to compare syntactic differences between two corpora. The method makes use of permutation tests to determine how reliable frequency differences between the corpora are, a technique that is gaining popularity in corpus linguistics (Lijffijt 2013). The first full dialectometric use of such an approach was by Sanders (2010), who used it to explore a Swedish dialect corpus. Let us exemplify the general idea starting with comparisons between county pairs.

Consider the following utterance from the Devon subcorpus of FRED-S:

(2) We$_{\texttt{PPIS2}}$ started$_{\texttt{VVD}}$ at$_{\texttt{II}}$ three$_{\texttt{MC}}$,$_{,}$ yes$_{\texttt{UH}}$.$_{.}$ <DEV_005>

Ignoring punctuation, we construct all overlapping sequences of length n (always $n = 2$ here, i.e, *bigrams*), yielding the following result: PPIS2_VVDI, VVD_II, II_MC, MC_UH. This is done for all utterances in both subcorpora, and the resulting bigram counts are aggregated on the county level. A normalization procedure that redistributes probability point mass is applied to the resulting absolute frequencies. More specifically, this normalization process keeps the total number of tokens per *n*-gram constant, but scales the per-county numbers according to the amount of *n*-grams in that county.[7]. Then, the data is randomly resampled (without replacement) based on turns[8] and the procedure is repeated on the new corpus. In other words, a new corpus is created, in which each county subcorpus contains the same amount of turns, but each turn is randomly assigned to a county. We can now, for each *n*-grams, calculate the difference in normalized values between the two counties both in the "true" (i.e. original) corpus and the randomly resampled one. If this difference is smaller in the original corpus, we can count this as evidence that the original difference may have occurred purely by chance - after all, a random process yielded a greater difference. Finding that the difference for the permuted corpus is smaller, however, would be consistent with the hypothesis that there is a genuine difference between the counties. If this process is not only done once, but a large number of times, the proportion of cases where the original difference was actually greater yields a probabilistic measure of how significant an *n*-grams frequency difference between the two

[7] This method was chosen based on the process described in Nerbonne & Wiersma (2006) The difference to a more familiar normalization scheme seems to be marginal – the correlation between raw distances derived using this method and those from simple per 10,000 *n*-grams normalization is greater than 0.99.

[8] Nerbonne & Wiersma (2006) resample based on sentences; in later work (Wiersma, Nerbonne & Lauttamus 2011), this was changed to resampling based on speakers to increase the reliability of the result. This was not feasible for this study, as the number of speakers for some of the counties was low (see also Sanders 2010).

counties is. In a similar fashion, the overall distance (computed using a suitable distance metric, such as the Manhattan distance) can be evaluated.

In addition, we can run the permutation not only on county pairs, but over all counties at the same time. Instead of comparing differences between counties, we calculate the proportion of random corpora in which the normalized frequency exceeds that of the true corpus, counting runs where they are equal as one half. This yields a *reliability matrix* that indicates how reliably this *n*-gram appears more (or less) frequently than expected by chance. Using either the pairwise number of significant differences or a measure of the extremeness of the distribution derived from the reliability matrix, we can evaluate particularly distinctive patterns. For example, consider `AT_NN1`, the definite article followed by a singular noun, the most frequent pattern. This pattern was significantly different in 84 of all the 136 possible pairwise county combinations; this amounts to being the 29th most distinct *n*-grams according to this metric, If, on the other hand, we use the number of divergences per *n*-gram from maximally consistent whole-corpus permutations, we obtain a value of 2.42 percent. This is still one of the top patterns, but ranks quite a bit lower at number 83 on the corresponding list. In general, a similar pattern seems to hold: there is a strong correlation between the two measures of distinctiveness, but the number of pairwise significant differences is more strongly influenced by *n*-grams frequency. Both the normalized frequency matrix and the reliability matrix are suitable for distance calculations. Details about the process can be found in Wolk (2014: 64-74).

Figure 4 displays the results. For comparison, the left-hand map displays the results of the normalization-based[9] top-down approach on the texts that are available in FRED-S. First, we note that the approach fares better here than on the complete data set: the clusters are spatially mostly coherent, geographic distance explains 27.6 percent of the variance in linguistic distances, and text size differences account for a slightly lower share of 48.5 percent. This results from the removal of most counties with particularly restricted textual coverage. The center map, showing a cluster analysis of normalized bottom-up frequencies, yields a much messier picture; moreover the quantitative match of the linguistic distances is considerably worse: the R^2 for geographic distances is much lower at 0.10, and that for text size considerably higher at a staggering 0.67. The right plot, finally, shows the clusters resulting from the reliability matrix, restricted to bigrams with at least seven pairwise significant differences. Here, the match between geographic and linguistic distances is almost as good as for

[9] This comparison is more appropriate than that to the model-based variant, as both do not employ geographic smoothing.

Figure 4: Cluster maps – hierarchical agglomerative cluster analysis using Ward's method as clustering algorithm (displayed: 3-cluster solution). Left: Top-down normalization-based approach on FRED-s. Center: Bottom-up normalized frequencies. Right: Bottom-up reliability scores.

the top-down result on the left at 26.2 percent, and the influence of corpus size is significantly lower ($R^2 = 0.16$). Qualitatively, we find that contiguous regions emerge. This analysis finds a division between Scotland and the North of England only at a position of lower importance, and instead emphasizes a transition area between the North and South of England. Finally, we add that both measures of distinctiveness identify dialectologically meaningful patterns. Among the ten most distinctive *n*-grams, for example, we find many known features of British dialect grammar: several bigrams related to *was/were* variation (PPH1/PPHS1/PPHS2_VBDR, *it/(he/she)/they were*, and PPHS2_VBDZ, *they was*), *them* as a determiner (in particular after temporal nouns, as in *them days*), *used to* as a marker of habituality (in particular following nouns), and *is n't/not*, which competes with the non-standard form *ain't*.

In short, then, our results suggest that bottom-up CBDM is practicable and worthwhile; the best method yielded results comparable to those for the manually selected feature set. The normalized frequencies, however, fared rather badly, and only after the permutation process smoothed off the rough edges did the areal signal emerge.

Finally, it bears mentioning that the nature of the tag set and the tagging procedure used have a strong influence on the linguistic patterns that emerge, and therefore most probably also on the aggregational results. Exploring the effect of changes in the data pipeline will be crucial in the further development of this approach.

5 Conclusion

In this contribution, we have sketched some recent advances in corpus-based dialectometry (CBDM). CBDM bases claims about geolinguistic patterns in aggregate linguistic variability on language usage as observed through naturalistic corpus data (as opposed to e.g. linguistic knowledge). Early studies in CBDM (e.g. Szmrecsanyi 2011) aggregate normalized text frequencies of features in an a-prioristically defined feature catalogue. This approach we have called the normalization-based top-down CBDM approach (see §3.1). What we have shown is that top-down CBDM profits from the probabilistic modeling of usage frequencies prior to aggregation. The other major advance we have dwelt on in this contribution is bottom-up CBDM, which does not draw on a pre-defined feature catalogue but lets the features to be aggregated emerge in a data-driven fashion through the identification of significant and/or distinctive part-of-speech *n*-grams. Both probabilistically enhanced top-down CBDM and bottom-up CBDM are valuable additions to the dialectometry toolbox.

With regard to the theme of the present volume, *The Future of Dialects*, we have discussed in this contribution new ways of analyzing dialect data. These new ways are in line with usage-based methodologies customary in related disciplines such as variationist sociolinguistics. As we have argued in the Introduction section, we believe that the future of dialectology will crucially include the ability to analyze actual usage data. As for the future of dialects *per se*, we would like to stress that a focus on more realistic, usage-oriented data sources is likely to reflect the many-faceted nature of dialects more faithfully than other methodologies do.

Acknowledgments

We wish to thank Peter Kleiweg for creating and maintaining the RuG/L04 package. The audience at the Workshop on 'Frontiers in the study of language variety' at the Methods XV conference in Groningen (August 2014) provided very helpful and valuable feedback on an earlier version of this paper. We also thank three anonymous reviewers for their extensive and constructive feedback. The second-named author gratefully acknowledges an Odysseus grant awarded by the Research Foundation Flanders (FWO, grant no. G.0C59.13N). The usual disclaimers apply.

References

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

Anderwald, Lieselotte & Benedikt Szmrecsanyi. 2009. Corpus linguistics and dialectology. In Anke Lüdeling & Merja Kytö (eds.), *Corpus linguistics. An international handbook* (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 29/1), 1126–1139. Berlin, New York: Mouton de Gruyter.

Bybee, Joan L. 2010. *Language, usage and cognition*. Cambridge, New York: Cambridge University Press.

Chambers, J. K. & Peter Trudgill. 1998. *Dialectology*. 2nd edn. Cambridge, New York: Cambridge University Press.

Garside, Roger & Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In Roger Garside, Geoffrey Leech & Tony McEnery (eds.), *Corpus annotation: Linguistic information from computer text corpora*, 102–121. London: Longman.

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

Goebl, Hans. 2005a. Dialektometrie. In Reinhard Köhler, Gabriel Altmann & Rajmund G. Piotrowski (eds.), *Quantitative linguistics / Quantitative Linguistik. An international handbook / Ein internationales Handbuch* (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science 27), 498–531. Berlin, New York: Walter de Gruyter.

Goebl, Hans. 2005b. La dialectométrie corrélative. Un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme. *Revue de Linguistique Romane* 69. 321–367.

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

Gooskens, Charlotte. 2005. Traveling time as a predictor of linguistic distance. *Dialectologia et Geolinguistica* 13. 38–62.

Grieve, Jack. 2009. *A corpus-based regional dialect survey of grammatical variation in written Standard American English*. Flagstaff: Northern Arizona University PhD dissertation.

Heeringa, Wilbert. 2004. *Measuring dialect pronunciation differences using Levenshtein distance*. Groningen: University of Groningen PhD Thesis.

Hernández, Nuria. 2006. *User's Guide to FRED*. Freiburg: University of Freiburg.

Jain, Anil K., M. Narasimha Murty & Patrick J. Flynn. 1999. Data clustering: A review. *ACM Computing Surveys* 31(3). 264–323.

Kilgarriff, Adam. 2001. Comparing corpora. *International Journal of Corpus Linguistics* 6(1). 97–133.

Lijffijt, Jefrey. 2013. *Computational methods for comparison and exploration of event sequences*. Espoo: Aalto University PhD thesis.

Nerbonne, John. 2013. How much does geography influence language variation? In Peter Auer, Martin Hilpert, Anja Stukenbrock & Benedikt Szmrecsanyi (eds.), *Space in language and linguistics: Geographical, interactional, and cognitive perspectives*, 220–236. Berlin, New York: Walter de Gruyter.

Nerbonne, John, Wilbert Heeringa & Peter Kleiweg. 1999. Edit distance and dialect proximity. In David Sankoff & Joseph Kruskal (eds.), *Time warps, string edits and macromolecules: The theory and practice of sequence comparison*, v–xv. Stanford: CSLI Press.

Nerbonne, John & Peter Kleiweg. 2007. Toward a dialectological yardstick. *Journal of Quantitative Linguistics* 14(2). 148–166.

Nerbonne, John & Christine Siedle. 2005. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72(2). 129–147.

Nerbonne, John & Wybo Wiersma. 2006. A measure of aggregate syntactic distance. In John Nerbonne & Erhard Hinrichs (eds.), *Linguistic distances. Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney, July, 2006*, 82–90.

Nerbonne, John, Peter Kleiweg, Franz Manni & Wilbert Heeringa. 2008. Projecting dialect differences to geography: Bootstrapping clustering vs. clustering with noise. In Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker (eds.), *Data analysis, machine learning, and applications. Proceedings of the 31st Annual Meeting of the German Classification Society*, 647–654. Berlin: Springer.

Rumpf, Jonas, Simon Pickl, Stephan Elspaß, Werner König & Volker Schmidt. 2009. Structural analysis of dialect maps using methods from spatial statistics. *Zeitschrift für Dialektologie und Linguistik* 76(3). 280–308.

Sanders, Nathan C. 2010. *A statistical method for syntactic dialectometry*. Bloomington: Indiana University PhD thesis.

Scherrer, Yves. 2012. Recovering dialect geography from an unaligned comparable corpus. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH* (EACL 2012), 63–71. Avignon, France: Association for Computational Linguistics.

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

Szmrecsanyi, Benedikt. 2013. *Grammatical variation in British English dialects: A study in corpus-based dialectometry*. Cambridge, New York: Cambridge University Press.

Szmrecsanyi, Benedikt & Nuria Hernández. 2007. *Manual of Information to accompany the Freiburg Corpus of English Dialects Sampler ("FRED-S")*. Freiburg: University of Freiburg.

Szmrecsanyi, Benedikt & Bernhard Wälchli (eds.). 2014. *Aggregating dialectology, typology, and register analysis: Linguistic variation in text and speech* (Lingua & litterae 28). Berlin: Walter de Gruyter.

Tagliamonte, Sali. 2012. *Variationist sociolinguistics: Change, observation, interpretation*. Malden, Oxford, Chichester: Wiley-Blackwell.

Tomasello, Michael. 2003. *Constructing a language: A usage-based theory of language acquisition*. Cambridge, Mass: Harvard University Press.

Trudgill, Peter. 1982. The contribution of sociolinguistics to dialectology. *Language Sciences* 4(2). 237–250.

Wieling, Martijn. 2012. *A quantitative approach to social and geographical dialect variation*. Groningen: University of Groningen PhD thesis.

Wiersma, Wybo, John Nerbonne & Timo Lauttamus. 2011. Automatically extracting typical syntactic differences from corpora. *Literary and Linguistic Computing* 26(1). 107–124.

Wolk, Christoph. 2014. *Integrating aggregational and probabilistic approaches to language variation*. Freiburg: University of Freiburg PhD thesis.