Analyzing the BBC *Voices* data: Contemporary English dialect areas and their characteristic lexical variants

[a]Martijn Wieling*, [b]Clive Upton and [b]Ann Thompson

[a] Department of Quantitative Linguistics, University of Tübingen, Germany; [b]School of English, University of Leeds, United Kingdom

*Corresponding author: Martijn Wieling, Department of Quantitative Linguistics, University of Tübingen, Wilhelmstraße 19, D-72074 Tübingen, wieling@gmail.com

Abstract

This study investigates data from the BBC *Voices* project, which contains a large amount of vernacular data collected by the BBC between 2004 and 2005. The project was designed primarily to collect information on vernacular speech around the United Kingdom for broadcasting purposes. As part of the project, a web-based questionnaire was created, to which tens of thousands of people supplied their way of denoting thirty-eight variables which were known to exhibit marked lexical variation. Along with their variants, those responding to the on-line prompts provided information on their age, gender, and —significantly for this study— their location, this being recorded by means of their postcode. In this study we focus on the relative frequency of the top-ten variants for all variables in every postcode area. By using hierarchical spectral partitioning of bipartite graphs, we are able to identify four contemporary geographical dialect areas together with their characteristic lexical variants. Even though these variants can be said to characterize their respective geographical area, they also occur in other areas, and not all people in a certain region use the characteristic variant. This supports the view that dialect regions are not clearly defined by strict borders, but are fuzzy at best.

Introduction

In 2004 and 2005, the British Broadcasting Corporation conducted a large-scale survey, BBC *Voices*, in order to obtain a contemporary view of English dialectal variation. People visiting a specially-constructed website were invited to offer their variants for thirty-eight variables that were known to exhibit marked lexical variation. Along with their lexical use, informants were asked to provide details of their age, gender, and geographical (post-coded) location and length of time resident there. Upwards of 75,000 people participated in this project to a greater or lesser degree, resulting in a substantial electronic dataset.

As dialectologists, we are interested in investigating geographical structure which might be present in our data. Given the large size of the *Voices* lexical dataset (more than 700,000 responses in total), we use quantitative methods from dialectometry to provide an aggregate view of the contemporary English dialectal landscape. Dialectometry originated in the 1970's (Séguy, 1973) to provide a more objective method of identifying dialect differences than by "cherry-picking" the features which support the analysis one wishes to settle on (Nerbonne, 2009).

Dialectometry has not been received very favorably by some traditional dialectologists, as aggregate analyses obscure the importance of individual linguistic features, on which they are required to focus for their often philologically-directed purposes. Consequently, there have been a number of attempts to develop quantitative methods which enable the identification of characteristic linguistic variables. For example, Shackleton (2007) uses cluster analysis and principal component analysis (PCA) to identify linguistic variables which show a specific geographic distribution, while Grieve et al. (2011) uses spatial autocorrelation to detect significant geographical patterns in forty individual lexical variables. Prokić et al. (2012) examine each item in a dataset, seeking those that differ minimally with a candidate area and maximally with respect to sites outside the area.

In this study we use hierarchical bipartite spectral graph partitioning (BiSGP: Dhillon, 2001), which allows the simultaneous identification of geographical areas and their characteristic linguistic features. This approach has been successfully used to obtain the linguistic basis (in terms of sound correspondences) with respect to a certain reference pronunciation for Dutch (Wieling et al., 2010), English (Wieling et al., forthcoming) and Tuscan (Montemagni et al., forthcoming) dialect datasets. In contrast to analyzing pronunciation data, however, we investigate the use of specific lexical variants from *Voices* data.

Dataset

The BBC *Voices* data contains a total of 38 variables (Table 1).

1. Hot	*2. Cold*	*3. Tired*	*4. Unwell*	*5. Pleased*
6. Annoyed	*7. Play a game*	*8. Play truant*	*9. Throw*	*10. Hit hard*
11. Sleep	*12. Drunk*	*13. Pregnant*	*14. Left-handed*	*15. Lacking money*
16. Rich	*17. Insane*	*18. Attractive*	*19. Unattractive*	*20. Moody*
21. Baby	*22. Mother*	*23. Grandmother*	*24. Grandfather*	*25. Friend*
26. Male partner	*27. Female partner*	*28. Young person in cheap trendy clothes and jewelry*	*29. Clothes*	*30. Trousers*
31. Child's soft shoes worn for PE	*32. Main room of house (with TV)*	*33. Long, soft seat in the main room*	*34. Toilet*	*35. Narrow walkway alongside buildings*
36. To rain lightly	*37. To rain heavily*	*38. Running water smaller than a river*		

Table 1: Variables in the BBC *Voices* dataset

The complete dataset contains (on average) 19,326 responses per variable. We include only responses from the online questionnaire, as the responses on the (identical) paper questionnaire have not been digitized. As a consequence of paper copies not being included, the average age of participants is relatively low (about 33) and more than sixty percent of the participants are aged below thirty. 57.3 percent of participants are female.

The responses were lemmatized in order to abstract away from variation in spelling. For example, <skive>, <scaive>, <scive> (for the variable PLAY TRUANT) were grouped together. To simplify the data somewhat, we only select the top ten variants for every variable (on average containing 84 percent of all responses). We group the responses by postcode area (there are 121 postcode areas in the UK) and for every (lemmatized) variant we calculate the percentage of people in the postcode area using it. Our input data thus consists of a table with 121 rows (the postcode areas) and 380 columns (38 variables having 10 variants each) containing these percentages.

Methods

Clustering postcode areas and their variants simultaneously

To cluster postcode areas and their variants *simultaneously*, we use hierarchical spectral partitioning of bipartite graphs (originally proposed by Dhillon, 2001 and first used in language variation studies by Wieling and Nerbonne, 2010). A bipartite graph is a graph having two sets of vertices (one representing postcode areas and the other variants per variable) and a set of edges connecting vertices from one set to the other set (each edge represents the occurrence of the variant in the postcode area). No other edges, for example between postcode areas, are allowed. Consequently, our input table (postcode areas × variants) can be taken as a representation of a bipartite graph. A percentage in a cell greater than zero indicates that there is an edge between a postcode area and a variant (and the percentage indicates the ‘thickness’ of the edge), while a value of zero indicates the absence of an edge. The bipartite spectral graph partitioning method is based on calculating the singular value decomposition of this input matrix. The hierarchical clustering is obtained by repeatedly clustering the input matrix into two groups. An extensive mathematical explanation as well as an example of the bipartite spectral graph partitioning method is provided by Wieling and Nerbonne (2010, 2011). It is important to note that while the final analysis shows variants within certain clusters and not in others, for any given variant this does not imply that that variant is not at all used outside the cluster in which it appears. Variants may (and are likely to) be used outside the cluster in which they appear, but they will be less used outside those clusters.

Determining the most important variants per cluster

As we have a large set of variants, any given cluster is likely to contain multiple variants. Of course, we are only interested in the most characteristic variants for every cluster. Wieling and Nerbonne (2011) propose a method to measure the importance of a linguistic feature (in our case a specific variant) in a cluster by combining two measures, representativeness and distinctiveness. Representativeness of a variant measures how frequently it occurs in the postcode areas in the cluster. For example, if a cluster consists of ten postcode areas and the variant occurs uniquely in six postcode areas, the representativeness is 0.6. Note that this measure is identical to Labov et al.’s (2006) homogeneity of an isogloss,

defined as the total hits for a variant within its isogloss divided by the total number of speakers within the isogloss. Distinctiveness of a variant measures how frequently the variant occurs within as opposed to outside the cluster (while taking the relative size of the clusters into account). For example, a distinctiveness of 1 indicates that the variant is not used outside of the cluster. If a cluster contains 50 percent of the postcode areas and 50 percent (or less) of the total variant occurrences, the distinctiveness is set to zero. Note that this measure is similar, but not identical to Labov et al.'s (2006) consistency of an isogloss, defined as the total number of hits for a variant within its isogloss divided by the total number of hits for that variant in the whole study. Our measure differs from theirs as we take the relative sizes of the areas into account and therefore correct for chance effects. The values of distinctiveness and representativeness range between zero and one. In order to prevent variants which are high in representativeness but very low in distinctiveness or vice versa obtaining a high ranking, we apply a minimum threshold of 0.1 for both measures. The final importance value for every variant is obtained by simply averaging the representativeness and distinctiveness.

As an example, consider the calculation of the representativeness and distinctiveness of the variant "ned" (for the variable YOUNG PERSON IN CHEAP TRENDY CLOTHES AND JEWELRY) in a cluster containing 14 (mostly Scottish) postcode areas. Within these 14 postcode areas, the average percentage of use of the variant "ned" is 69%. Consequently, the representativeness (or Labov et al.'s homogeneity) equals 0.69. Given that 73% of all occurrences of "ned" in the dataset are located in this cluster, Labov et al.'s consistency measure equals 0.73. However, our distinctiveness measure takes the relative size of the cluster in account, by subtracting the relative size from the consistency value and then dividing this value by one minus the relative size. As the relative size equals 14 divided by 121 (0.116), the distinctiveness equals (0.73 – 0.116) / (1 – 0.116) = 0.69. Finally, the importance of "ned" in the cluster is determined by averaging the representativeness and distinctiveness values, resulting in a value of 0.69.

To see that using the consistency measure of Labov et al. (2006) might be problematic, consider an increased cluster size of 88 postcode areas (73%) instead of the original 14 postcode areas. Obviously Labov et al.’s consistency measure remains the same, as it is size independent. Our distinctiveness value, however, would reduce to (0.73 – 0.73) / (1 – 0.73) = 0. In this case the variant is not distinctive, as it occurs as frequently in the cluster as would be expected on the basis of chance (i.e. 73% of the occurrences of “ned” are found in 73% of all postcode areas). Clearly, the use of distinctiveness is beneficial over the use of Labov et al.’s (2006) consistency.

Results

We use the hierarchical bipartite spectral graph partitioning method to obtain a clustering into four groups (i.e. two times a clustering in two groups). While the clustering procedure we employ is able to yield an arbitrarily high number of clusters, initial analyses —using several one-dimensional clustering algorithms and noisy clustering available in the online application Gabmap (Nerbonne et al., 2011)— revealed that only four clusters could reliably be obtained. Figure 1 shows the geographical distribution of the four clusters. Note that the first partitioning into two groups separated Scotland (regions 1 and 2 in Figure 1) from the rest of Great Britain (regions 3 and 4 in Figure 1). The second round of partitioning separated the northern part of Scotland (region 1) from the southern part of Scotland (region 2), while also separating the middle part of the U.K. (region 3) from the remaining varieties (region 4). Results generally match those of traditional UK dialectology (Trudgill, 1999).

Going through the clusters from north to south, Figures 2 to 5 show the five most characteristic lexical variants for every cluster (if there are five). Numbers in parentheses refer to the variable-identifying numbers used in Table 1. The top-most cluster (marked with number 1 in Figure 1) only consists of the postcode areas Lerwick and Kirkwall, and this cluster is characterized by just four lemmatized word variants (the remaining variants did not reach the threshold for representativeness and distinctiveness) of which only the first is linguistically interesting as the others are simply identical to the label of the variable: “bonnie” to denote the variable ATTRACTIVE, “cold” (instead of e.g., “freezing”) to denote the variable COLD, “rain” (instead of e.g., “drizzle”) to denote the variable TO RAIN LIGHTLY, and “grandfather” (instead of e.g., “granddad”) to denote the variable GRANDFATHER. It is clear that no variant has perfect distinctiveness, and that variants which are characteristic of cluster 1 also occur outside that cluster. As a consequence, cluster 1 is not perfectly characterized by a single variant but rather by its lack of distinction.

The situation is somewhat different for the remaining part of Scotland (indicated by number 2 in Figure 1). All of the five variants indicated have a very high distinctiveness as they generally (but never exclusively) occur in the top-most cluster. Note that “wean” also occurs frequently in Northern Ireland, but this is unsurprising given the Scotland-Ireland connection through Ulster Scots. The most characteristic variants are “ned” to denote the variable YOUNG PERSON IN CHEAP TRENDY CLOTHES AND JEWELRY, “burn” to denote the variable RUNNING WATER SMALLER THAN A RIVER, “dog (off)” to denote the variable PLAY TRUANT, “wean” to denote the variable BABY, and “close” to denote the variable NARROW WALKWAY ALONGSIDE BUILDINGS. The last three variants do have a high distinctiveness, but they have a lower representativeness since they do not occur in all postcode areas in the cluster.

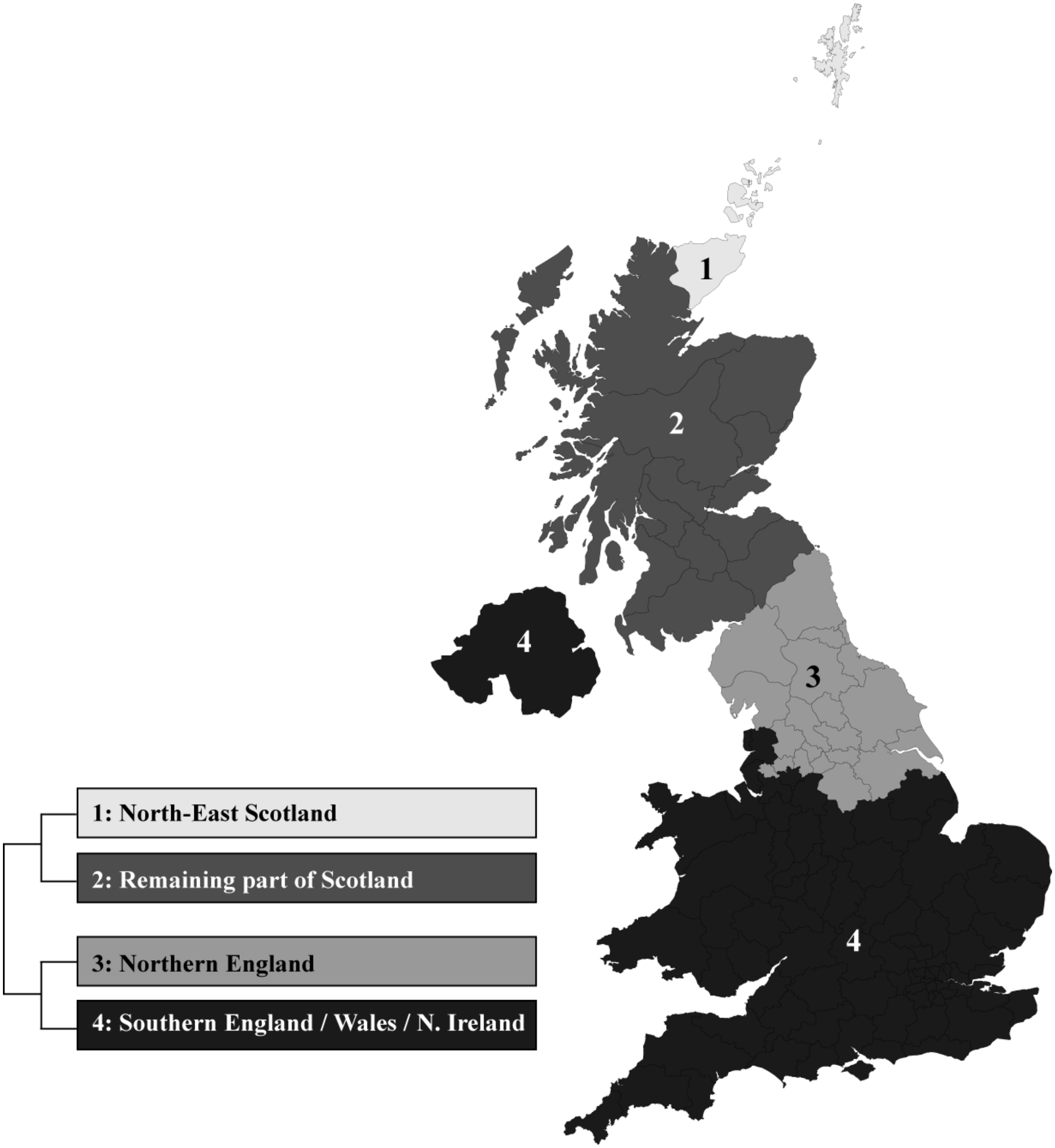
1
2
4
3
4
1: North-East Scotland
2: Remaining part of Scotland
3: Northern England
4: Southern England / Wales / N. Ireland

Figure 1. The four main clusters revealed by the bipartite spectral graph partitioning. The first split separates Scotland from England, Wales, and Northern Ireland dialects. The second split subdivides each of these areas.

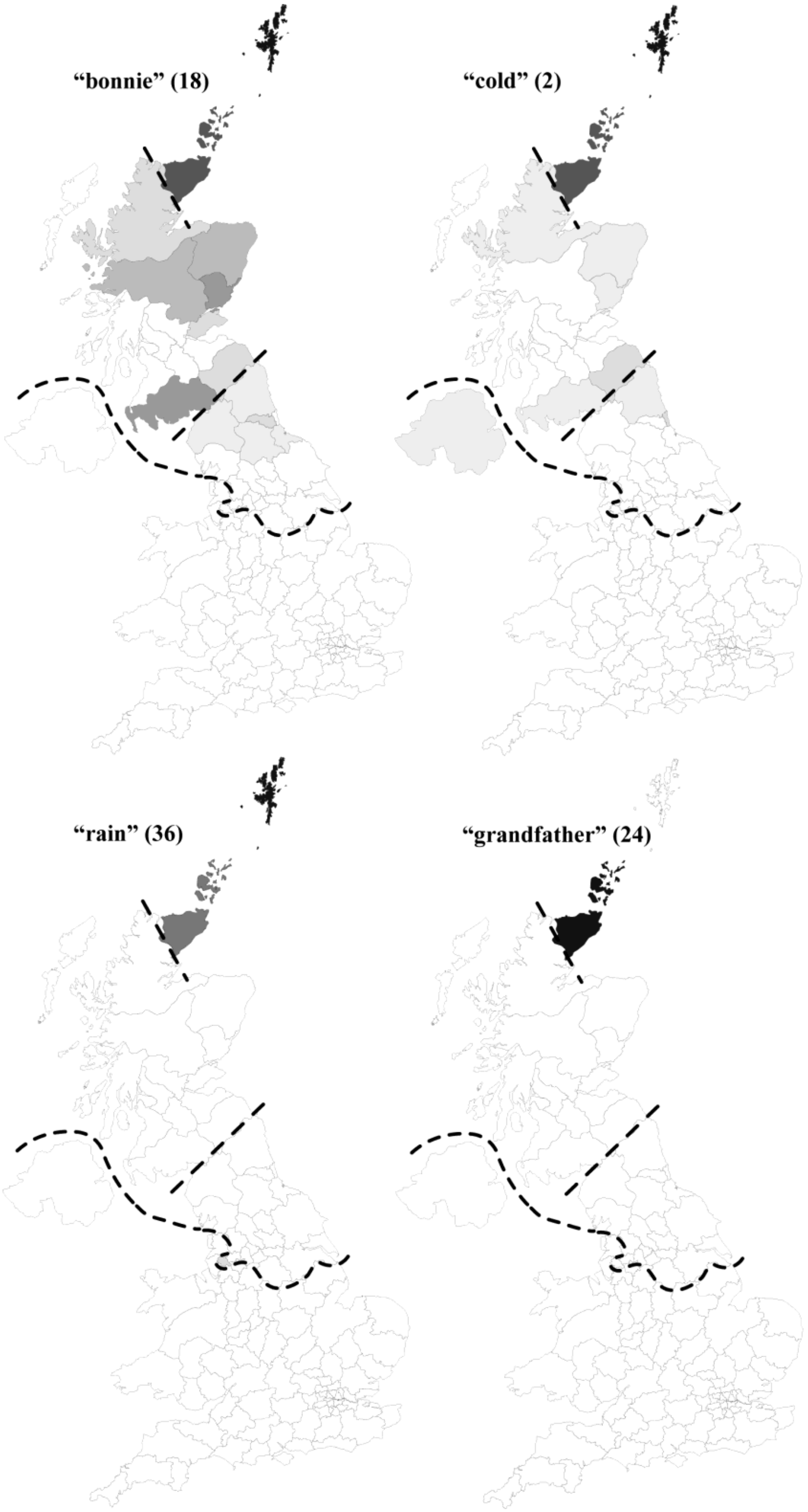

Figure 2. Most important variants for the top-most cluster (marked with number 1 in Figure 1). Darker shades of gray indicate a higher frequency of occurrence. Note that there are only four characteristic variants. The cluster borders are marked by dashed lines.

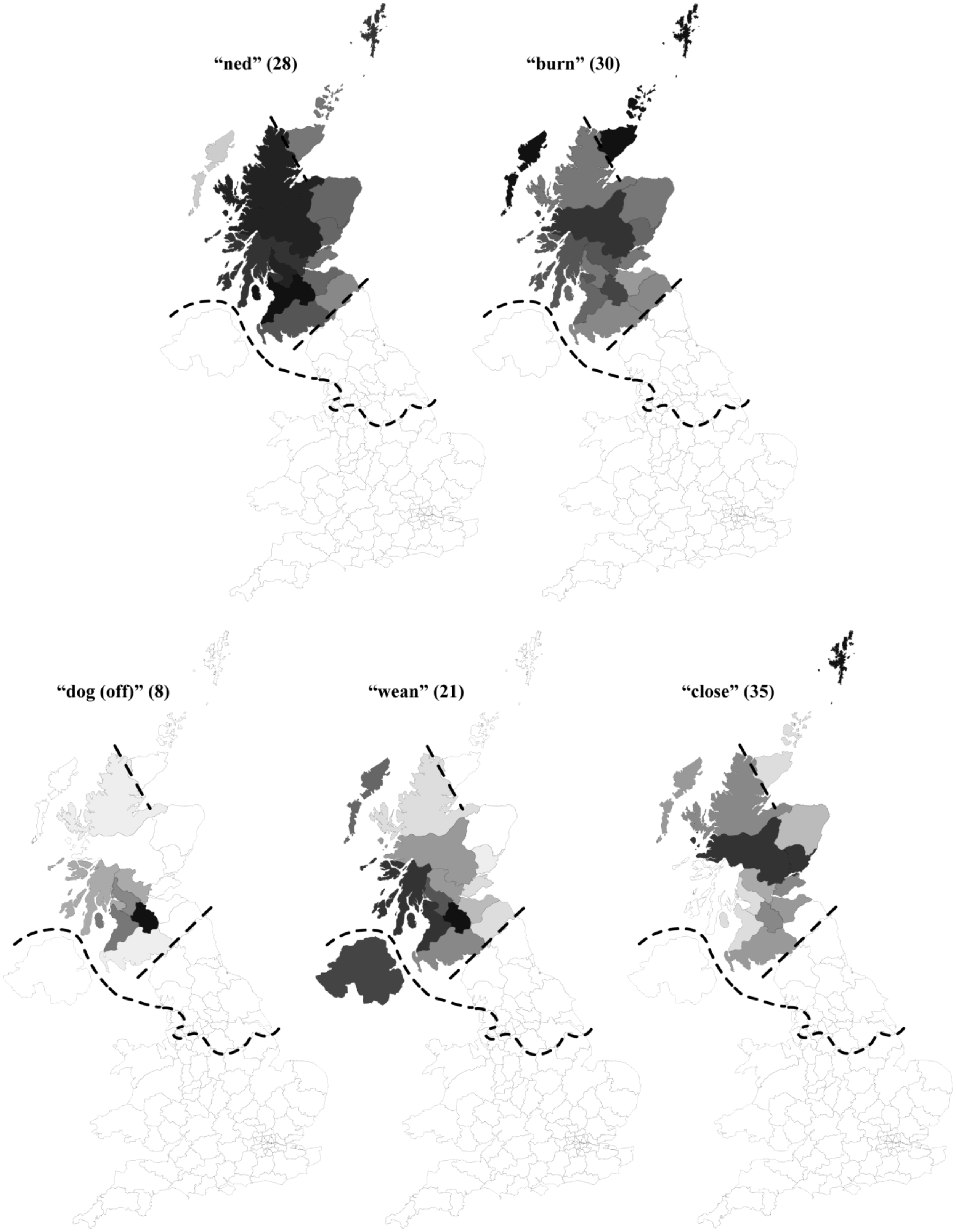

Figure 3. Most important variants for the Scottish cluster (marked with number 2 in Figure 1). Darker shades of gray indicate a higher frequency of occurrence. The cluster borders are marked by dashed lines.

The central cluster (marked by number 3 in Figure 1) is defined by relatively distinctive variants, but these variants are less representative as they never occur in all postcode areas of the cluster. The most characteristic variants are “beck” to denote the variable RUNNING WATER SMALLER THAN A RIVER, “lass” to denote the variable FEMALE PARTNER, “laik” to denote the variable PLAY A GAME, “ginnel” to denote the variable NARROW WALKWAY ALONGSIDE BUILDINGS, and “hoy” to denote the variable THROW.

The distinctiveness-representativeness pattern found for the central cluster is inverted for the southern cluster (marked by number 4 in Figure 1). The characteristic variants (except for “daps” and “plimsolls”) occur in most of the postcode areas in the cluster, but they also occur relatively frequently outside the cluster. Consequently, the representativeness is relatively high, while the distinctiveness is relatively low. The most characteristic variants are “stream” to denote the variable RUNNING WATER SMALLER THAN A RIVER, “alley” to denote the variable NARROW WALKWAY ALONGSIDE BUILDINGS, “daps” (which only occurs in the Southwest) and “plimsolls” (which mainly occurs in the Southeast) to denote the variable CHILD’S SOFT SHOES WORN FOR PHYSICAL EDUCATION, and “chav” to denote the variable YOUNG PERSON IN CHEAP TRENDY CLOTHES AND JEWELRY.

It is clear that the clusters are not separated from each other by clear-cut borders in the sense that there is a specific variant or combination of variants only occurring in one cluster (and in all the postcode areas in one cluster) but never in another. Consequently, the dialect area boundaries are fuzzy.

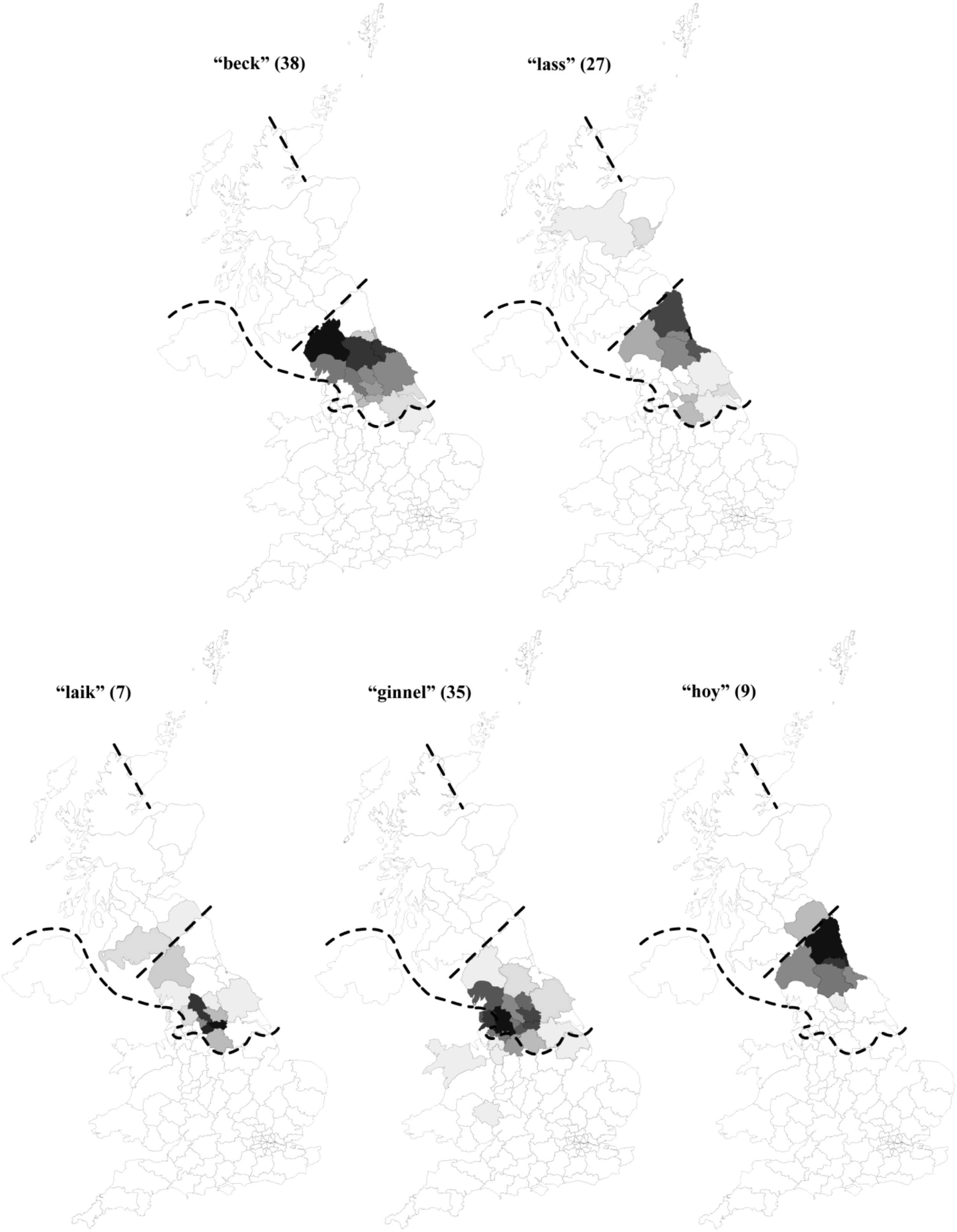

Figure 4. Most important variants for the central cluster (marked by number 3 in Figure 1). Darker shades of gray indicate a higher frequency of occurrence. The cluster borders are marked by dashed lines.

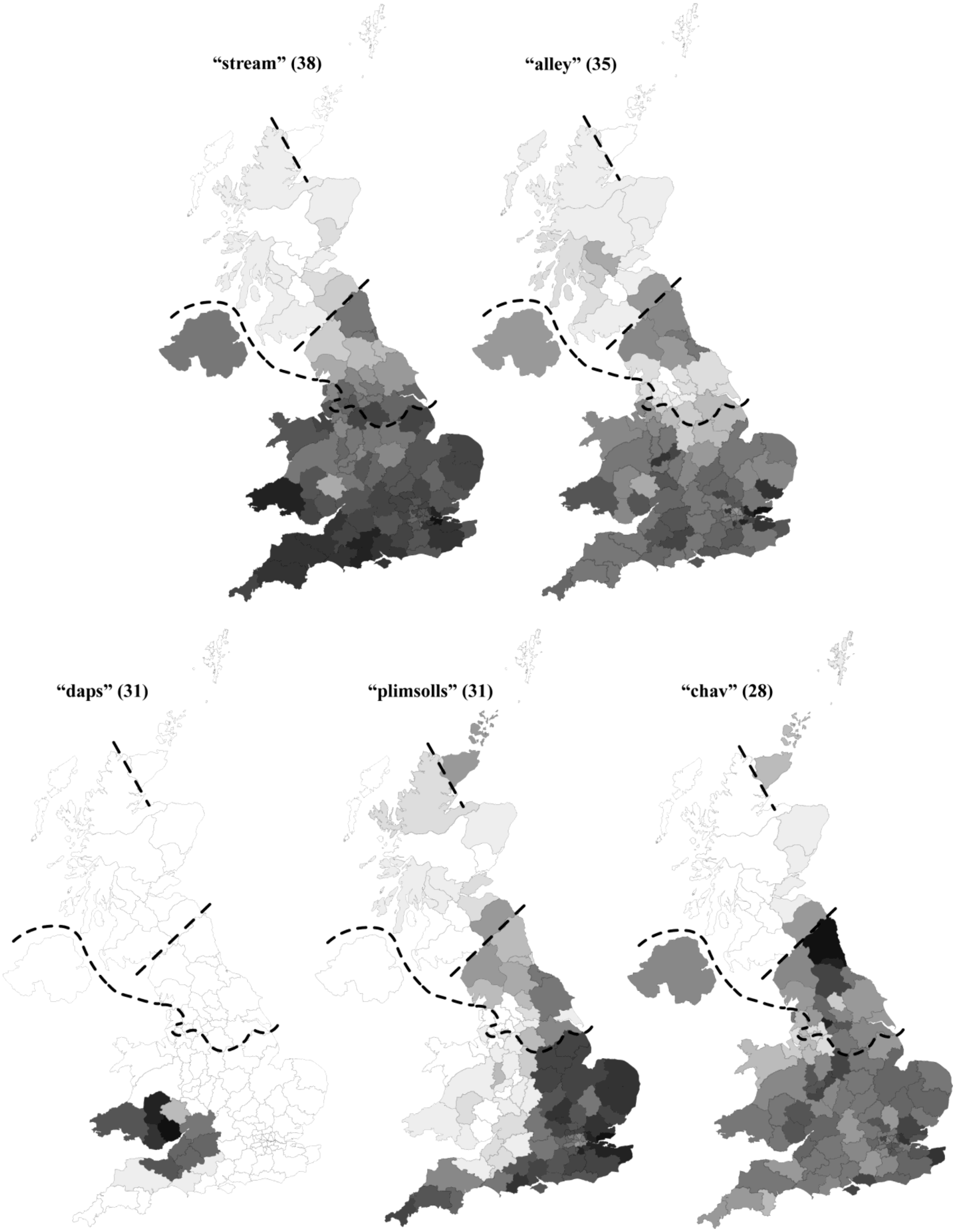

Figure 5. Most important variants for the southern cluster (marked by number 4 in Figure 1). Darker shades of gray indicate a higher frequency of occurrence. The cluster borders are marked by dashed lines.

Discussion

In this study we have shown that bipartite spectral graph partitioning can be usefully employed to identify significant dialectal areas for contemporary English, in particular its lexical variation. The distribution of variants also illustrates that there are no clear borders between the dialect areas (although Scotland is distinguished relatively clearly). Characteristic variants for one cluster can appear in another, and no two variants emerging from the analysis as individually distinctive exhibit precisely the same distributions. Distinctiveness of a whole area is thus essentially a relative rather than an absolute attribute.

Beyond the essential fuzziness that surrounds the demarcation of areas, what is apparent is that the comparatively distinctive variants emerging in areas 2 and 3 are without exception non-standard in terms of *English* Standard English: “ned”, “burn”, “dog off”, “wean”, “close”, “beck”, “lass”, “laik”, “ginnel” and “hoy” will all readily be thought non-standard by most native English speakers, many of whom will be able to identify at least some of them with the areas identified on the maps. (See for example Wright, 1898-1905 and Upton et al., 1994 for distributions of many of these variants.) This is not to say, of course, that those distinctive of area 2, especially “burn”, “wean”, and “close” (along with “bonnie” from area 1), are actually essentially non-standard, as they are well recognized as standard forms in *Scottish* English and so have a status as such. The fact that all five of the area-2 variants occur only north of the England-Scotland border marks them out as distinctly Scottish. We can contrast the confined distributions of the variants in areas 2 (Scotland excluding its North-East) and 3 (Northern England) to the situation for those in areas 1 (North East Scotland) and 4 (Southern England and Northern Ireland). Here, characteristic words which emerge from the analysis are in the main *English* Standard English ones. The exceptions here are noteworthy, and contrasting. “Bonnie” and “wean” are words which are widely associated with Scots and Scottish English but which are also found further south, in Northern England They are representative rather than distinctive, like the other items in areas 1 and 4 which have still wider Standard currency. “Daps”, by contrast, is noticeably extremely localized to the English Southwest and to South Wales, highly distinctive of these areas (and high in the consciousness of its users as such).

Therefore, as is especially apparent in areas 2 and 3 but is also made clear in one instance in area 4, distinctiveness in dialectal spatial differentiation is associated with non-standard lexis. In contrast, the use of lexis widely considered standard is predictably shown to be characteristic of greater diffusion and so notably less of distinctiveness, whilst at the same time (with the exception of “bonnie”) being significantly of an essentially southern-English concentration.

Of course, there are still outstanding issues which might be investigated. Due to the large size of the dataset, we opted only to investigate the top ten variants per variable. While this makes sense from an aggregate perspective, the approach might exclude some variants occurring in only a few postcode areas. Consequently, especially for the smallest cluster, we might miss some characteristic variants. In addition, while our lemmatization step grouped together many words which can be seen as the same variant, in some cases it is not immediately apparent if two words should be grouped or separated.

References

Dhillon, I. (2001). Co-clustering documents and words using bipartite spectral graph partitioning. *Proceedings of the seventh Association for Computing Machinery Special Interest Group on Knowledge Discovery and Data Mining (ACM SIGKDD).* NY: ACM, 269-274.

Grieve, J., Speelman, D. and Geeraerts, D. (2011). A statistical method for the identification and aggregation of regional linguistic variation. *Language Variation and Change* 23: 193-221.

Labov, W., Ash, S. and Boberg, C. (2006). *Atlas of North American English: Phonology and Phonetics*. Berlin: Mouton de Gruyter.

Montemagni, S., Wieling, M., de Jonge, B. and Nerbonne, J. (forthcoming). Patterns of language variation and underlying linguistic features: a new dialectometric approach. In: Nicola de Blasi (ed.) *La variazione nell'italiano e nella sua storia. Varietà e varianti linguistiche e testuali. Proceedings of the Congresso della Società Internazionale di Linguistica e Filologia Italiana (XI Congresso SILFI).*

Nerbonne, J. (2009). Data-Driven Dialectology, *Language and Linguistics Compass* 3: 175-198.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P. and Leinonen, T. (2011). Gabmap — A Web Application for Dialectology. *Dialectologia*. Special Issue II: 65-89.

Orton, H. and Dieth, E. (1962). *Survey of English Dialects*. Leeds: E.J. Arnold.

Prokić, J., Çöltekin, Ç. and Nerbonne, J. (2012). Detecting Shibboleths. In: Miriam Butt and Jelena Prokić (eds.) *Visualization of Language Patterns and Uncovering Language History from Multilingual Resources.* Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France: Association for Computational Linguistics.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de Gascogne, *Revue de Linguistique Romane*, 37: 1–24.

Shackleton, R. G., Jr. (2007). Phonetic variation in the traditional English dialects: a computational analysis, *Journal of English Linguistics*, 33: 99-160.

Trudgill, P. (1999). *The Dialects of England*, 2nd revised edition. Oxford: Blackwell.

Upton, C, Parry D. and Widdowson J.D.A. (1994). *Survey of English Dialects: The Dictionary and Grammar*. London: Routledge.

Wieling, M. and Nerbonne, J. (2010). Hierarchical spectral partitioning of bipartite graphs to cluster dialects and identify distinguishing features. In Banea, C., Moschitti, A., Somasundaran, S. and Zanzotto, F.M. (eds.), *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*. Stroudsburg, PA: Association for Computational Linguistics, pp. 33-41.

Wieling, M. and Nerbonne, J. (2011). Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features, *Computer Speech and Language*, 25: 700-715.

Wieling, M., Shackleton, Jr., R.G. and Nerbonne, J. (forthcoming). Analyzing phonetic variation in the traditional English dialects: Simultaneously clustering dialect and phonetic features. *LLC: The Journal of Digital Scholarship in the Humanities.*

Wright, J. ed. (1898-1905). *The English Dialect Dictionary*. Oxford: Clarendon Press.