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

*Martijn Wieling*,  *Clive Upton* and  *Ann Thompson*

*Center for Language and Cognition, University of Groningen, The Netherlands; School of English, University of Leeds, United Kingdom

*Corresponding author: Clive Upton, School of English, University of Leeds, Leeds, LS2 9JT, United Kingdom,
c.s.upton@leeds.ac.uk
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. This 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 concepts which were known to exhibit marked lexical variation. Along with their variants, those responding to the online 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 concepts 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 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 concepts 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. Upwards of 29,000 people participated in this project (“BBC Voices”) to a greater or lesser degree, resulting in a substantial electronic dataset as a consequence.

As dialectologists we are interested in investigating geographical structure which might be present in our data. Given the enormous size of the *Voices* lexical dataset (containing 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).

Unfortunately, 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 alternation 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, however, we use hierarchical bipartite spectral graph partitioning (BiSGP), which allows a simultaneous identification of geographical areas together with 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., submitted) and Tuscan (Montemagni et al., forthcoming) dialect datasets. In contrast to analyzing pronunciation data, however, we investigate the use of specific lexical variants (per concept variable) from *Voices* data.

**Dataset**

The BBC *Voices* data contains a total of 38 concepts which are shown in Table 1 below.

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31. Child’s soft shoes worn for PE</td>
<td>32. Main room of house (with TV)</td>
<td>33. Long, soft seat in the main room</td>
<td>34. Toilet</td>
<td>35. Narrow walkway alongside buildings</td>
</tr>
<tr>
<td>36. To rain lightly</td>
<td>37. To rain heavily</td>
<td>38. Running water smaller than a river</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: List of all 38 concepts in the BBC Voices dataset
The complete dataset contains (on average) 19,326 responses per concept. We only include 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 the people is relatively low (about 33) and more than sixty percent of the people were aged below thirty. A total of 57.3 percent of the participants were female.

The responses were lemmatized in order to abstract away from variation in spelling. For example, ‘skive’, ‘scaive’, ‘scive’ (for the concept PLAY TRUANT) were grouped together. To simplify the data somewhat, we only select the top-ten variants for every concept (on average containing 84 percent of all responses). We group the responses by postcode area (there are a total of 121 UK postcodes) and for every (lemmatized) variant we calculate the percentage of people in the postcode area using this variant. Our input data thus consists of a table with 121 rows (the postcode areas) and 380 columns (38 concepts 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 (Wieling and Nerbonne, 2010). A bipartite graph is a graph having two sets of vertices (one representing postcode areas and the other variants per concept) 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 in 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).

\textit{Determining the most important variants per cluster}

As we have a large set of variants, many will be assigned to a particular cluster. 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). 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 frequency, the distinctiveness is set to zero. 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.

**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 in two groups separated Scotland from the rest of Great Britain.

Going through the clusters from north to south, Figures 2 to 5 show the five (if present) most characteristic lexical variants for every cluster. Numbers in parentheses refer to the concept-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): “bonnie” to denote the concept ATTRACTIVE, “cold” (instead of e.g., “freezing”) to denote the concept COLD, “rain” (instead of e.g., “drizzle”) to denote the concept TO RAIN LIGHTLY, and “grandfather” (instead of e.g., “granddad”) to denote the concept GRANDFATHER. It is clear that no variant has perfect distinctiveness and characteristic variants also occur outside the cluster. As a consequence, this cluster is not perfectly characterized by a single variant.
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 concept YOUNG PERSON IN CHEAP TRENDY CLOTHES AND JEWELRY, “burn” to denote the concept RUNNING WATER SMALLER THAN A RIVER, “dog (off)” to denote the concept PLAY TRUANT, “wean” to denote the concept BABY, and “close” to denote the concept 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.

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 concept RUNNING WATER SMALLER THAN A RIVER, “lass” to denote the concept FEMALE PARTNER, “laik” to denote the concept PLAY A GAME, “ginnel” to denote the concept NARROW WALKWAY ALONGSIDE BUILDINGS, and “hoy” to denote the concept 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 of the cluster. Consequently, the representativeness is relatively high, while the distinctiveness is relatively low. The most characteristic variants are “stream” to denote the concept RUNNING WATER SMALLER THAN A RIVER, “alley” to denote the
concept NARROW WALKWAY ALONGSIDE BUILDINGS, “daps” (which only occurs in the southwest) and “plimsolls” (which mainly occurs in the southeast) to denote the concept CHILD’S SOFT SHOES WORN FOR PHYSICAL EDUCATION, and “chav” to denote the concept 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 are specific words 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 1. The four main clusters revealed by the bipartite spectral graph partitioning. The first split separates northern (Scottish) from southern (English and Northern Ireland) dialects. A 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.
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.
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.
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.
Discussion

In this study we have shown that bipartite spectral graph partitioning can be usefully employed to identify significant dialectal areas on the basis of contemporary English, in particular its lexical variation. The distribution of variants of the concept variables also illustrates that there are no clear borders between the dialect areas. Characteristic variants for one cluster can appear in another and no two features 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 indentified 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. And the confining of all five of the area-2 variants in these data north of the England-Scotland border marks them out as distinctly Scottish. We can contrast the confined distributions of the variants in areas 2 and 3 to the situation for those in areas 1 and 4. Here, characteristic words which emerge from the analysis are in the main English Standard English ones. The exceptions here are noteworthy, and contrasting. “Bonnie” is a word which is widely associated with Scots and Scottish English but which is found further south into northern England also: it is 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.

Beyond this confirmation of intuitions, outstanding issues remain 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, this approach might exclude some variants only occurring in 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


