Detecting Shibboleths

Jelena Prokić
Ludwig-Maximilians-Universität
j.prokic@lmu.de

Çağrı Çöltekin
University of Groningen
c.coltekin@rug.nl

John Nerbonne
University of Groningen
j.nerbonne@rug.nl

Abstract

A SHIBBOLETH is a pronunciation, or, more generally, a variant of speech that betrays where a speaker is from (*Judges* 12:6). We propose a generalization of the well-known precision and recall scores to deal with the case of detecting distinctive, characteristic variants when the analysis is based on numerical difference scores. We also compare our proposal to Fisher's linear discriminant, and we demonstrate its effectiveness on Dutch and German dialect data. It is a general method that can be applied both in synchronic and diachronic linguistics that involve automatic classification of linguistic entities.

1 Introduction and Background

The background of this contribution is the line of work known as DIALECTOMETRY (Séguy, 1973; Goebl, 1982), which has made computational work popular in dialectology. The basic idea of dialectometry is simple: one acquires large samples of corresponding material (e.g., a list of lexical choices, such as the word for carbonated soft drink, which might be 'soda', 'pop', 'tonic' etc.) from different sites within a language area, and then, for each pair of samples, one counts (or more generally measures) the difference at each point of correspondence. The differences are summed, and, given representative and sufficiently large samples, the results characterizes the degree to which one site differs from another.

Earlier work in dialectology mapped the distributions of individual items, recording lines of division on maps, so-called ISOGLOSSES, and then sought bundles of these as tell-tale indicators of important divisions between DIALECT AREAS. But as Chambers & Trudgill (1998) note, the earlier methodology is fraught with problems, many of which stem from the freedom of choice with respect to isoglosses, and their (normal) failure to 'bundle' neatly. Nerbonne (2009) notes that dialectometry improves on the traditional techniques in many ways, most of which stem from the fact that it shifts focus to AGGREGATE LEVEL of differences. Dialectometry uses large amounts of material; it reduces the subjectivity inherent in choosing isoglosses; it frequently analyzes material in ways unintended by those who designed dialect data collection efforts, including more sources of differences; and finally it replaces search for categorical overlap by a statistical analysis of differences.

Dialectometry does not enjoy overwhelming popularity in dialectology, however, and one of the reasons is simply that dialectologists, but also laymen, are interested not only in the aggregate relations among sites, or even the determination of dialect areas (or the structure of other geographic influence on language variation, such as dialect continua), but are quite enamored of the details involved. Dialectology scholars, but also laymen, wish to now where 'coffee' is ordered (in English) with a labialized /k/ sound ([k^{w}ɔfi]) or where in Germany one is likely to hear [p] and where [p͡f] in words such as *Pfad* 'path' or *Pfund* 'pound'.

Such characteristic features are known as SHIBBOLETHS, following a famous story in the old testament where people were killed because of where they were from, which was betrayed by their inability to pronounce the initial [ʃ] in the word 'shibboleth' (*Judges* 12:6). We propose a generalization of the well-known precision and

Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH, pages 72–80,
Avignon, France, April 23 - 24 2012. ©2012 Association for Computational Linguistics

recall scores, appropriate when dealing with distances, and which are designed to detect distinctive, characteristic variants when the analysis is based on numerical difference scores. We also compare our proposal to Fisher's linear discriminant, and we demonstrate its effectiveness on Dutch and German dialect data. Finally we evaluate the success of the proposal by visually examining an MDS plot showing the distances one obtains when the analysis is restricted to the features determined to be characteristic.

The paper proceeds from a dialectometric perspective, but the technique proposed does not assume an aggregate analysis, only that a group of sites has been identified somehow or another. The task is then to identify characteristic features of (candidate) dialect areas.

1.1 Related Work

Wieling and Nerbonne (2011) introduced two measures seeking to identify elements characteristic of a given group, REPRESENTATIVENESS and DISTINCTIVENESS. The intuition behind representativeness is simply that a feature increases in representativeness to the degree that it is found at each site in the group. We simplify their definition slightly as they focus on sound correspondences, i.e. categorical variables, while we shall formulate ideas about features in general.

$$\text{Representativeness}(f, g) = \frac{|g^f|}{|g|}$$

where f is a feature (in their case sound correspondence) in question, g is the set of sites in a given cluster, and g^f denotes the set of sites where feature f is observed.

As Wieling (2012) notes, if one construes the sites in the given group as 'relevant documents' and features as 'queries', then this definition is equivalent to RECALL in information retrieval (IR).

The intuition behind distinctiveness is similar to that behind IR's PRECISION, which measures the fraction of positive query responses that identify relevant documents. In our case this would be the fraction of those sites instantiating a feature that are indeed in the group we seek to characterize. In the case of groups of sites in dialectological analysis, however, we are dealing with groups that may make up significant fractions of the entire set of sites. Wieling and Nerbonne therefore introduced a correction for 'chance instantiation'. This is derived from the relative size of the group in question:

$$\begin{aligned}
\text{RelSize}(g) &= \tfrac{|g|}{|G|} \\
\text{RelOcc}(f, g) &= \tfrac{|g^f|}{|G^f|} \\
\text{Distinct}(f, g) &= \tfrac{\text{RelOcc}(f,g) - \text{RelSize}(g)}{1 - \text{RelSize}(g)}
\end{aligned}$$

where, G is the set of sites in the larger area of interest.

As a consequence, smaller clusters are given larger scores than clusters that contain many objects. Distinctiveness may even fall below zero, but these will be very uninteresting cases — those which occur relatively more frequently outside the group under consideration than within it.

Critique

There are two major problems with the earlier formulation which we seek to solve in this paper. First, the formulation, if taken strictly, applies only to individual values of categorical features, not to the features themselves. Second, many dialectological analyses are based on numerical measures of feature differences, e.g., the edit distance between two pronunciation transcriptions or the distance in formant space between two vowel pronunciations (Leinonen, 2010).

We seek a more general manner of detecting characteristic features below, i.e. one that applies to features, and not just to their (categorical) values and, in particular, one that can work hand in hand with numerical measures of feature differences.

2 Characteristic Features

Since dialectometry is built on measuring differences, we assume this in our formulation, and we seek those features which differ little within the group in question and a great deal outside that group. We focus on the setting where we examine one candidate group at a time, seeking features which characterize it best in distinction to elements outside the group.

We assume therefore, as earlier, a group g that we are examining consisting of $|g|$ sites among a larger area of interest G with $|G|$ sites including the sites s both within and outside g. We further explicitly assume a measure of difference d between sites, always with respect to a given feature

f. Then we calculate a mean difference with respect to f within the group in question:

$$\bar{d}^{g}_{f} = \frac{2}{|g|^2 - |g|} \sum_{s,s' \in g} d_f(s, s')$$

and a mean difference with respect f involving elements from outside the group:

$$\bar{d}^{\not g}_{f} = \frac{1}{|g|(|G| - |g|)} \sum_{s \in g, s' \notin g} d_f(s, s')$$

We then propose to identify characteristic features as those with relatively large differences between $\bar{d}^{\not g}_{f}$ and $\bar{d}^{g}_{f}$. However, we note that scale of these calculations are sensitive to a number of factors, including the size of the group and the number of individual differences calculated (which may vary due to missing values). To remedy the difficulties of comparing different features, and possibly very different distributions, we standardize both $\bar{d}^{\not g}_{f}$ and $\bar{d}^{g}_{f}$ and calculate the difference between the *z-score*s, where mean and standard deviation of the difference values are estimated from all distance values calculated with respect to feature f. As a result, we use the measure

$$\frac{\bar{d}^{\not g}_{f} - \bar{d}_f}{sd(d_f)} - \frac{\bar{d}^{g}_{f} - \bar{d}_f}{sd(d_f)}$$

where d_f represents all distance values with respect to feature f (the formula is not simplified for the sake of clarity). We emphasize that we normalized the difference scores for each feature separately. Had we normalized with respect to *all* the differences, we would only have transformed the original problem in a linear fashion.

Note that this formulation allows us to apply the definitions to both categorical and to numerical data, assuming only that the difference measure is numerical. See illustration in Figure 1.

For this work we used a difference function that finds the aggregated minimum Levenshtein distance between two sites as calculated by Gabmap (Nerbonne et al., 2011). However, we again emphasize that the benefit of this method in comparison to others proposed earlier is that it can be used with any feature type as long as one can define a numerical distance metric between the features. Regardless of the type of data set, some distance values between certain sites may not be possible to calculate, typically due to missing values. This may affect the scale and the reliability of the average distance calculations presented above. For the experiments reported below, we calculated average scores only if the missing values did not exceed 20% of the total values used in the calculation.

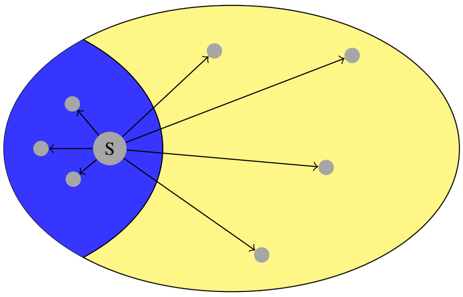


Figure 1: Illustration of the calculation of a distance function. Our proposal compares the mean distance of all pairs of sites within a group, including all those shown on the left (in blue) to the mean distance of the pairs of sites where the first is within the group and the second outside it.

Fisher's Linear Discriminant

The formulation we propose looks a good deal like the well-known Fisher's linear discriminant (FLD) (Schalkoff, 1992, 90ff), which maximizes the differences in means between two data sets with respect to (the sum of) their variances.

$$S = \frac{\sigma^2_{between}}{\sigma^2_{within}}$$

But FLD is defined for vectors, while we wish to generalize to cases where only *differences* are guaranteed to be numerical measures. The mean of categorical features, for example, is undefined. We might imagine applying something like FLD in the space of differences, but note that low variance does not necessarily correspond to a tightly knit group in difference space. If we measure the differences among all the pairs of sites in a candidate group, each of which realizes a given categorical feature differently, the mean difference of pairs will be one (unit) and the variance zero. Difference spaces are simply constructed differently.

Silhouette method

We also note relation of our approach to the SILHOUETTE method introduced by Rousseeuw (1987) used to evaluate clustering validity. The silhouette method is used to determine the optimal number of clusters for a given dataset. It starts from data that has already been clustered using any of the (hierarchical or flat) clustering techniques. For every object i in the data (these would be sites in clustering to detect dialect groups) it calculates the average dissimilarity to all other objects in the same cluster $a(i)$, and the average dissimilarity to all objects in all other clusters (for every cluster separately). After the distances to all other clusters are computed, the cluster with the smallest average distance ($b(i)$) to the object in question is selected as the most appropriate one for that object. The silhouette $s(i)$ is calculated as

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$

Values close to 1 indicate that the object is appropriately clustered, while negative values indicate that the object should have been clustered in its neighbouring cluster. By comparing silhouette values obtained by clustering into different numbers of groups, this technique indicates an optimal clustering.

We compare average distances within groups to average distance to objects outside groups with respect to individual features, making our proposal different. A second point of difference is that we aim not to score 'groupings', but rather how characteristic specific features are for a given grouping.

3 Experimental set up

The method we propose is tested on Dutch and German dialect data. We use Levenshtein algorithm in order to calculate the distances between the sites and Ward's clustering method to group the sites. In this section we give a brief description of the data and the clustering procedure.

Dutch data set

Dutch dialect data comes form the Goeman-Taeldeman-Van Reenen Project[1] that comprises 1876 items collected from more than 600 locations in the Netherlands and Flanders. The data was collected during the period 1979-1996, transcribed into IPA and later digitalized. It consists of inflected and uninflected words, word groups and short sentences. More on this project can be found in Goeman and Taeldeman (1996).

The data used in this paper is a subset of the GTRP data set and consist of the pronunciations of 562 words collected at 613 location in the Netherlands and Flanders. It includes only single word items that show phonetic variation. Multi-word items and items that show morphological, rather than phonetic variation, were excluded from the analysis. Items where multiple lexemes per site are possible were also excluded.[2]

German data set

German dialect data comes from the project 'Kleiner Deutscher Lautatlas — Phonetik' at the 'Forschungszentrum Deutscher Sprachatlas' in Marburg. In this project a number of sentences from Georg Wenker's huge collection of German dialects (1870s-1880s)[3] were recorded and transcribed in the late 1970s and early 1990s (Göschel, 1992). The aim of the project was to give an overview of the sound structure of modern German dialects.

In this paper we use a small subset of the data that consists of the transcriptions of 40 words. We have selected only words that are present at all or almost all 186 locations evenly distributed over Germany.

Distance matrices

The distances between each pair of sites within each of the two data sets were calculated using the Levenshtein algorithm (Levenshtein, 1966). This method is frequently used in dialect comparison to measure the differences between two sites (Nerbonne et al., 1996; Heeringa, 2004). It aligns two strings and calculates the number of mismatching segments in two strings. The total distance between two sites is the average distance between all compared strings collected at those two sites. For the method proposed in this paper, any other method whose output is a numerical distance metric between the features can be applied. The final result is a *site* × *site* distance matrix, that can later be analyzed by means of clustering or, alternatively, using a dimensionality reduction technique such multidimensional scaling.

We analyze two distance matrices using Ward's clustering algorithm, also known as the minimal variance algorithm. We use *MDS plots* (as implemented in Gabmap (Nerbonne et al., 2011)) as a visual basis to choose the optimal number for clusters for the two data sets. The choice of the

[1] http://www.meertens.knaw.nl

[2] The data set used in this paper can be downloaded from http://www.gabmap.nl/ app/examples/.

[3] See, too, the *Digitaler Wenker Atlas* (DiWA) project, http://www.3.diwa.info/

appropriate clustering algorithm is a difficult task as is the determination of the number of significant groups (Prokić and Nerbonne, 2008), but these questions are not the subjects of this paper. At the risk of repeating ourselves, we emphasize that our focus in this paper is not the choice of clustering method or the determination of the most significant (number of) groups. We do not even assume that the groups were obtained via clustering, only that candidate groups have somehow been identified. We focus then on finding the most characteristic features for a given group of sites. In the next section we present the results of applying our method to the Dutch and German data sets.

Evaluation

We evaluate success in the task of selecting items characteristic of an area by using MDS to analyze a distance matrix obtained from only that item. We then project the first, most important MDS dimension to a map asking whether the original group of sites indeed is identified. Note that in successful cases the area corresponding to the group may be shaded either as darker than the rest or as lighter. In either case the item (word) has served to characterize the region and the sites in it.

We also experimented with clustering to analyze the distances based on the pronunciations of the candidate characteristic shibboleths, but single word distances unsurprisingly yielded very unstable results. For that reason we use MDS.

4 Results

Dutch

We examine a clustering of the distance matrix for Dutch varieties with six clusters, which we present in Figure 2.

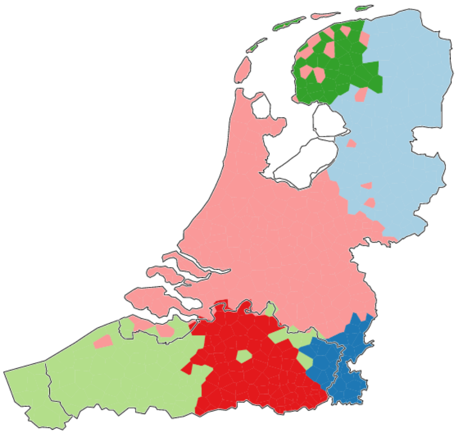


Figure 2: Six dialect groups in Dutch speaking area.

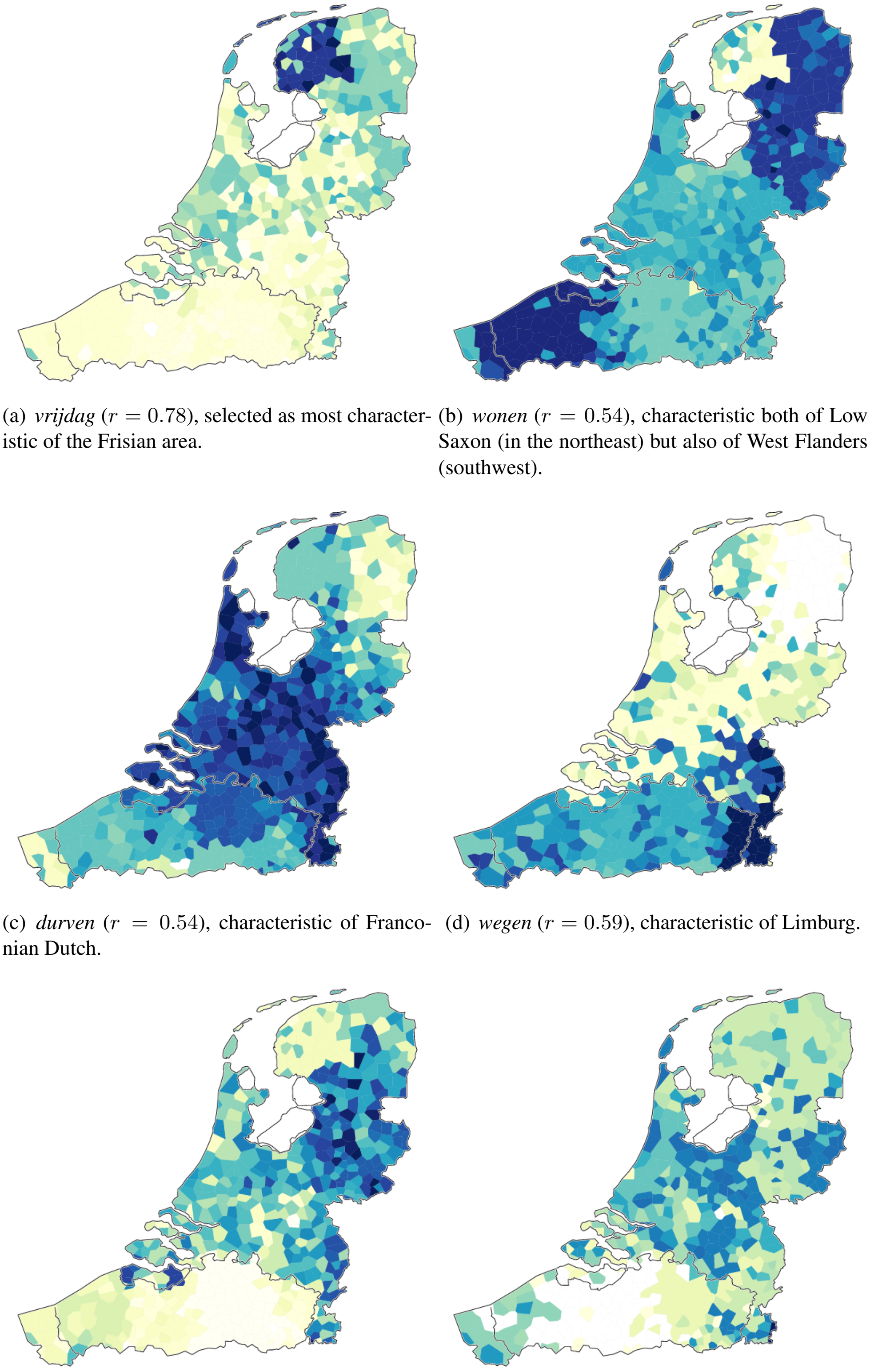
The clustering algorithm identified Frisian (dark green), Low Saxon (Groningen and Overijsel, light blue), Dutch Franconian varieties (pink), Limburg (dark blue), Belgian Brabant (red) and West Flanders (light green) dialect groups. For each feature (word) in our data set and for each group of sites (cluster) we calculated the differences within the given site and also with respect to each of the other five groups in order to determine which words differ the least within the given group and still differ a great deal with respect to the sites outside the group. The top five words for each group of sites are presented in Table 1.

The results obtained show that the same word could be prominent for more than one cluster; for example, the word *scheiden* is scored highly in two different dialect groups. In Figure 3 we present maps of Dutch language area that are based on the pronunciations of the best scoring words for each of the six groups of sites. For each word we calculated the Levenshtein distance and analyzed the resulting distance matrices using MDS. In maps in Figure 3 we present the first extracted dimension, which always explains most of the variation in the data.[4] We also supply the degree to which the extracted dimensions correlate with the distances in the input matrix.

Maps in Figure 3 reveal that the best scoring word does indeed identify the cluster in question. For example, the map in Figure 3(a) reveals that based on the pronunciation of word *vrijdag* the Frisian-speaking area is internally homogeneous and distinct from the rest of the sites. No other groups can be identified in the map. In Figure 3(b) we present the analysis of a distance matrix based on the pronunciation of the word *wonen* ‘live’ that was found to be relevant for the Low Saxon area. The map shows two areas, Low Saxon and West Flanders, where it was also among top 10 best scored words, as two distinct areas.[5]

[4] The only exception is Figure 3(b) where we present second dimension.

[5] These two areas are both known for pronouncing the *slot ’n* in final unstressed syllables of the form /ən/ as a syllabic nasal that has assimilated in place to the preceding consonant.



(a) *vrijdag* ($r = 0.78$), selected as most characteristic of the Frisian area.

(b) *wonen* ($r = 0.54$), characteristic both of Low Saxon (in the northeast) but also of West Flanders (southwest).

(c) *durven* ($r = 0.54$), characteristic of Franconian Dutch.

(d) *wegen* ($r = 0.59$), characteristic of Limburg.

(e) *gisteren* ($r = 0.60$), selected as characteristic of Belgian Brabant.

(f) *heet* ($r = 0.58$), selected as characteristic of West Flanders, but in fact not awfully successful in distinguishing exactly that area.

Figure 3: Dutch dialect area based on the pronunciation of words (a) *vrijdag*, (b) *wonen*, (c) *durven*, (d) *wegen*, (f) *heet* and (e) *gisteren* selected as characteristic of respective areas.

Frisian		Low Saxon		Franconian		Limburg		West Flanders		Belg.Brabant	
2.891217	vrijdag	1.881354	wonen	1.131973	durven	2.317413	wegen	1.605255	heet	1.968656	gisteren
2.808631	zoet	1.875302	dopen	1.101160	maanden	2.048480	schoenen	1.587253	weten	1.803535	gewoon
2.659577	geven	1.784224	scheiden	1.096989	metselen	2.015069	schaven	1.573224	weer	1.794680	gal
2.618426	draden	1.747136	bijten	1.073387	houden	1.979678	schapen	1.567049	keuren	1.764176	kleden
2.606748	dun	1.721321	worden	1.054981	dorsen	1.956787	scheiden	1.548940	horen	1.753901	wippen

Table 1: Five most characteristic words for each Dutch dialect variety.

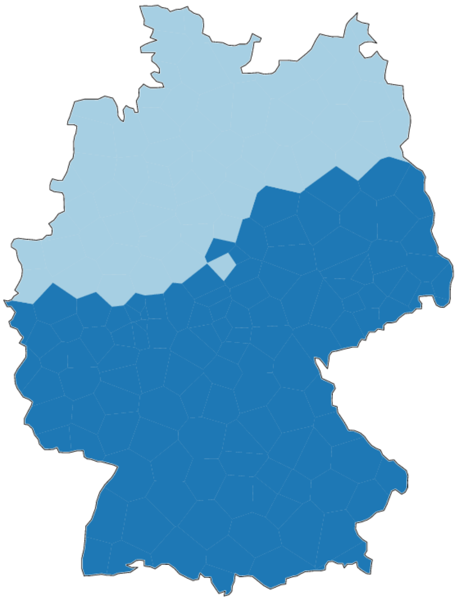


Figure 4: Two dialect groups in Germany.

German

We ran the same analysis for the German data set. In Figure 4 we present the two largest groups in the cluster analysis of the distances obtained using 40 words. We might have examined more groups, but we wished to examine results based on larger groups as well.

We focus on the top-level, two-way split that divides Germany into north and south.[6] These areas correspond with the traditional division into Low German on one hand, and Middle and High German on the other. Just as with the Dutch data, for every word in the data set and for each group of sites we calculate the distances with respect to the word in order to see how well the words characterize one of the two dialect groups. The results are presented in Table 2. Because we are examining a two-way split, it is not surprising that the same words sometimes characterize the areas (inversely).

North		South	
1.057400	weisse	1.056600	gefahre
1.011804	gefahre	0.909610	gross
0.982128	bleib	0.825211	weisse
0.920354	Ochse	0.764463	Pfeffer
0.831812	gross	0.755694	baue

Table 2: Five most prominent words for two dialect groups in Germany. Because we examine a two-way split, some words characterize both areas.

In Figures 5(a) and 5(b) we present the MDS maps based on the distances derived from comparing the words *weisse* and *gefahre*, which were two best ranked words.

The word *weisse* shows only small differences within the north, which is illustrated by the light-colored northern part of Germany in Figure 5(a). The map in Figure 5(b) shows an even clearer split highlighting the High German area based on the best ranked word found by our method. This word shows also low variation in the Low German area (second best scored), which is also clearly visible in Figure 5(b).

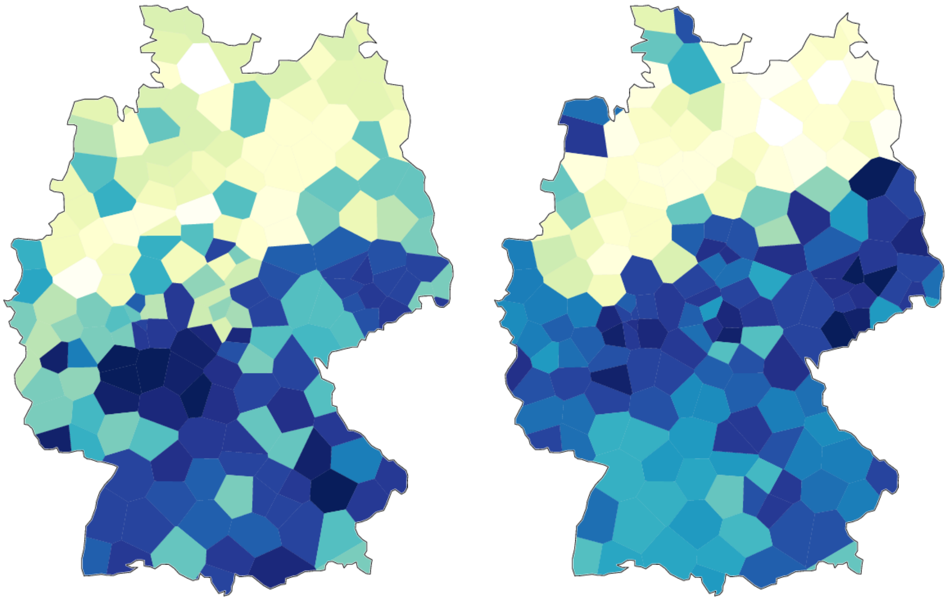
5 Conclusions

In this paper we have presented a method to detect the most characteristic features of a candidate group of linguistic varieties. The group might be one obtained from cluster analysis, but it might also be obtained from correspondence analysis (Cichocki, 2006), or it might simply be another group identified for theoretical or extra-linguistic reasons (geography or social properties).

The method is applicable to any feature type as long as one can define a numerical distance metric between the elements. In particular the method maybe applied to categorical data whose differences are individually zero or one, or to vowels characterized by the Euclidean distance between formant vectors (or pairs), and it may be applied to edit distance measures applied to phonetic transcriptions. The proposed method is therefore not constrained in its application to only the categorical features, as the proposal in Wieling & Nerbonne (2011) was.

Essentially the method seeks items that differ minimally within a group but differ a great deal

[6] In anticipation of worries about the analysis we hasten to add that more finely discriminated groups may also be distinguished. That is not our purpose here.



(a) *weisse* ($r = 0.63$)

(b) *gefahre* ($r = 0.59$)

Figure 5: First MDS dimensions based on the pronunciation of words (a) *weisse* and (b) *gefahre*.

with respect to elements outside it. We crucially limited its application to elements that were instantiated at least 20% of the sites, and we used normalized z-scores in order to improve the comparability of the measurements.

We demonstrated the effectiveness of the proposed method on real dialect data by trying to identify the words that show low variation within a given dialect area, and high variation outside a given area. We evaluate the results of these experiments by visually examining the distances induced from single words. Although this indicated that the technique is performing well, we concede that alternative evaluations would be worth while, e.g. simply mapping the density of low distances between pairs in the distance matrix. This awaits future work.

The proposed method can be used in dialectometry to automatically identify characteristic features in dialect variation, while at the same time it offers traditional dialectologists insights into the details involved. Its application may also not be limited to dialectology (including dialectometry). It is a general method that can be applied in other branches of linguistics, such as historical linguistics or typology, that deal with language classification at various levels.

The method proposed in this paper might also find use in the evaluation of clustering, specifically in helping researchers to determine the optimal number of groups in a clustering solution. It might then result in a modification of the silhouette technique discussed earlier.

Application of computational methods in dialectology and historical linguistics is still not generally accepted. This state of affairs is due less to the questions that the groups of researchers are trying to answer, and more to the methods they are using to reach their goals. Bringing them together is a challenging task. The method we propose can analyse large amounts of data without losing sight of the linguistic details.

References

J.K. Chambers and Peter Trudgill. 1998. *Dialectology*. Cambridge University Press, Cambridge.

Wladyslaw Cichocki. 2006. Geographic variation in Acadian French /r/: What can correspondence analysis contribute? *Literary and Linguistic Computing*, 21(4):529–542. Special Issue, J.Nerbonne & W.Kretzschmar (eds.), *Progress in Dialectometry: Toward Explanation*.

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

Antonie Goeman and Johan Taeldeman. 1996. Fonologie en morfologie van de nederlandse dialecten. een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval*, 48:38–59.

Joachim Göschel. 1992. Das Forschungsinstitut für Deutsche Sprache "Deutscher Sprachatlas". Wissenschaftlicher Bericht, Das Forschungsinstitut für Deutsche Sprache, Marburg.

Wilbert Heeringa. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Ph.D. thesis, Rijksuniversiteit Groningen.

Therese Leinonen. 2010. *An Acoustic Analysis of Vowel Pronunciation in Swedish Dialects*. Ph.D. thesis, University of Groningen.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting insertions, deletions and reversals. *Cybernetics and Control Theory*, 10(8):707–710. Russian orig. in *Doklady Akademii Nauk SSR* 163(4), 845–848, 1965.

John Nerbonne, Wilbert Heeringa, Erik van den Hout, Peter van der Kooi, Simone Otten, and Willem van de Vis. 1996. Phonetic distance between dutch dialects. In Gert Durieux, Walter Daelemans, and Steven Gillis, editors, *CLIN VI: Proc. from the Sixth CLIN Meeting*, pages 185–202. Center for Dutch Language and Speech, University of Antwerpen (UIA), Antwerpen.

John Nerbonne, Rinke Coleand, Charlotte Gooskens, Peter Kleiweg, and Therese Leinonen. 2011. Gabmap: A web application for dialectology. *Dialectologia*, Special issue II:65–89.

John Nerbonne. 2009. Data-driven dialectology. *Language and Linguistics Compass*, 3(1):175–198.

Jelena Prokić and John Nerbonne. 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing*, 2(1-2):153–172. DOI: 10.13366/E1753854809000366.

Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Robert Schalkoff. 1992. *Pattern Recognition: Statistical, Structural and Neural Approaches*. John Wiley, New York.

Jean Séguy. 1973. La dialectométrie dans l'atlas linguistique de gascogne. *Revue de Linguistique Romane*, 37(145):1–24.

Martijn Wieling and John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language*, 25:700–715. DOI:10.1016/j.csl.2010.05.004. Published online May 21, 2010.

Martijn Wieling. 2012. *A Quantitative Approach to Social and Geogrpahical Dialect Variation*. Ph.D. thesis, University of Groningen.