Using Gabmap

Therese Leinonen[1], Çağrı Çöltekin[2], and John Nerbonne[2]

[1]University of Turku

[2]University of Groningen

February 2, 2015

Abstract

Gabmap is a freely available, open-source web application that analyzes the data of language variation, e.g. varying words for the same concepts, varying pronunciations for the same words, or varying frequencies of syntactic constructions in transcribed conversations. Gabmap is an integrated part of CLARIN (see e.g. http://portal.clarin.nl).[1]. This article summarizes Gabmap's basic functionality, adding material on some new features and reporting on the range of uses to which Gabmap has been put. Gabmap is modestly successful, and its popularity underscores the fact that the study of language variation has crossed a watershed concerning the acceptability of automated language analysis. Automated analysis not only improves researchers' efficiency, it also improves the replicability of their analyses and allows them to focus on inferences to be drawn from analyses and other more abstract aspects of that study.

1 Introduction

Gabmap is a freely available, open-source web application that analyzes the data of language variation, e.g. varying words for the same concepts, varying pronunciations for the same words, or varying frequencies of syntactic constructions in transcribed conversations.

Other possibilities exist as well, but these are by far the most frequent uses to which Gabmap has been put. Nerbonne et al. (2011) reports on Gabmap's basic functionality and its implementation, so that this article can build on that, adding material on new functionality and reporting on the range of uses to which Gabmap has been put. Gabmap is modestly

[1]We are grateful to CLARIN-NL and for their support of the project ADEPT (http://www.clarin.nl/node/70#ADEPT), which in turn produced Gabmap. CLARIN-NL was supported by the Netherlands Organization for Scientific Research (NWO)

successful, and its popularity underscores the fact that the study of language variation has crossed a watershed concerning the acceptability of automated language analysis. Automated analysis not only improves researchers' efficiency, it also improves the replicability of their analyses and allows them to focus on inferences to be drawn from analyses and other more abstract aspects of that study.

2 A Gabmap session

In this section, we show an example of a typical Gabmap session and the types of analyses that can be conducted. For this purpose we use data from the Goeman-Taeldeman-Van Reenen-project (GTRP; Goeman and Taeldeman 1996). The data consist of phonetic transcriptions of Dutch dialects from the Netherlands and Belgium gathered during the period 1980—1995. These data are available as demo data on the Gabmap web site, which makes it possible for users to try out the analyses described here directly in Gabmap.

2.1 Data

The dialect data can be prepared in a spreadsheet where rows represent sites and columns represent linguistic variables. In the demo data, the columns are words and each cell in the spreadsheet shows the pronunciation of a word in the International Phonetic Alphabet (IPA) at one specific site:[2]

	boter	**broden**	**zout**
Aalsmeer	botər	brojə	zaut
Baardegem	botər	bruəs	zat
Coevorden	bœtər	brodn	sɔl̞t

Gabmap accepts tab-separated Unicode text files as input data, and most spreadsheet software allow exporting data to text files with Unicode encoding.

Analysis in Gabmap is not restricted to transcribed pronunciation data; instead, any kind of binary or numeric data can be used. When uploading data into Gabmap, the type of data is specified, so that the data can be processed appropriately. For the phonetic transcriptions in the example we choose *string data* as the type of data and *string edit distance* as the type of processing (more about data processing in section 2.3).

In order to create dialect maps, the data file should be accompanied by a map file with the geographical coordinates of the data sites and optionally borders of the country

[2] If there are several pronunciations available of a word from one site, these can be separated by "space slash space" in the data file.

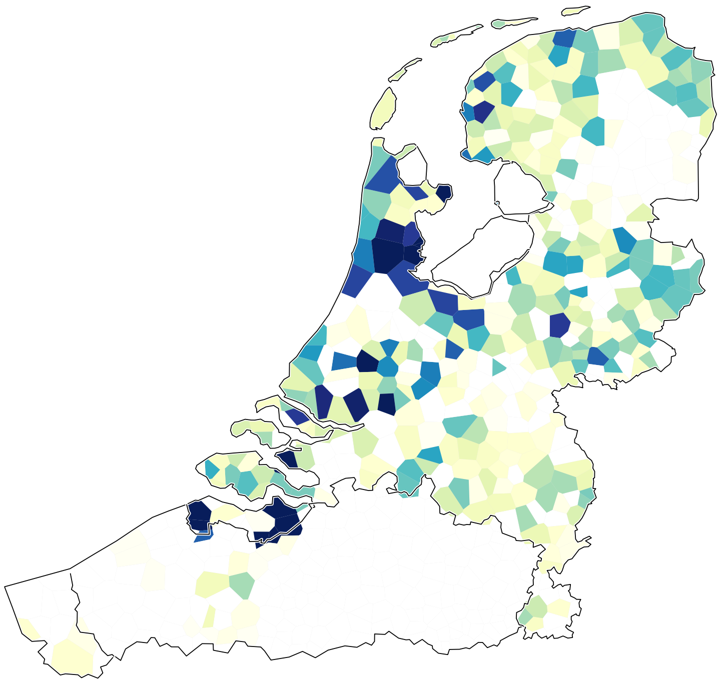

Figure 1. Distribution map of the character [ɫ].

or language area. The map file is a .kml or .kmz file that can be created in Google Earth or using the Google Maps service through any standard web browser. Using a map file is, however, not compulsory. Users might want to analyze language variation related to other factors than geography. The data rows might, for example, be individual speakers instead of sites. For analysis of this type of data, no map file is needed and Gabmap will create a pseudo map instead of real maps in the mapping functions. The statistical analyses, like cluster analysis and multidimensional scaling (see below), will, then, show how individual speakers group together based on their language use.

When a project is created, Gabmap offers several ways of inspecting the data. Summaries are created of the number of sites, number of words (or other linguistic variables), number of characters and number of tokens. In *Data overview* in Gabmap, we can, for example, see that the demo data has data from 613 places and that the number of different words (items) is 562. The total number of word transcriptions (instances) is 331,690, which is less than 613×562 due to some missing data in the input table.

2.2 Distribution maps

Several types of distribution maps are offered in Gabmap. Figure 1 shows a map of one specific phonetic character in the data set. The character maps are part of the data overview function in Gabmap, where maps can be created of any character or token in the data set. Figure 1 shows the distribution of the velarized lateral approximant [ɫ]. White color means no instances at all of the character from a site, and the darker the color the higher the

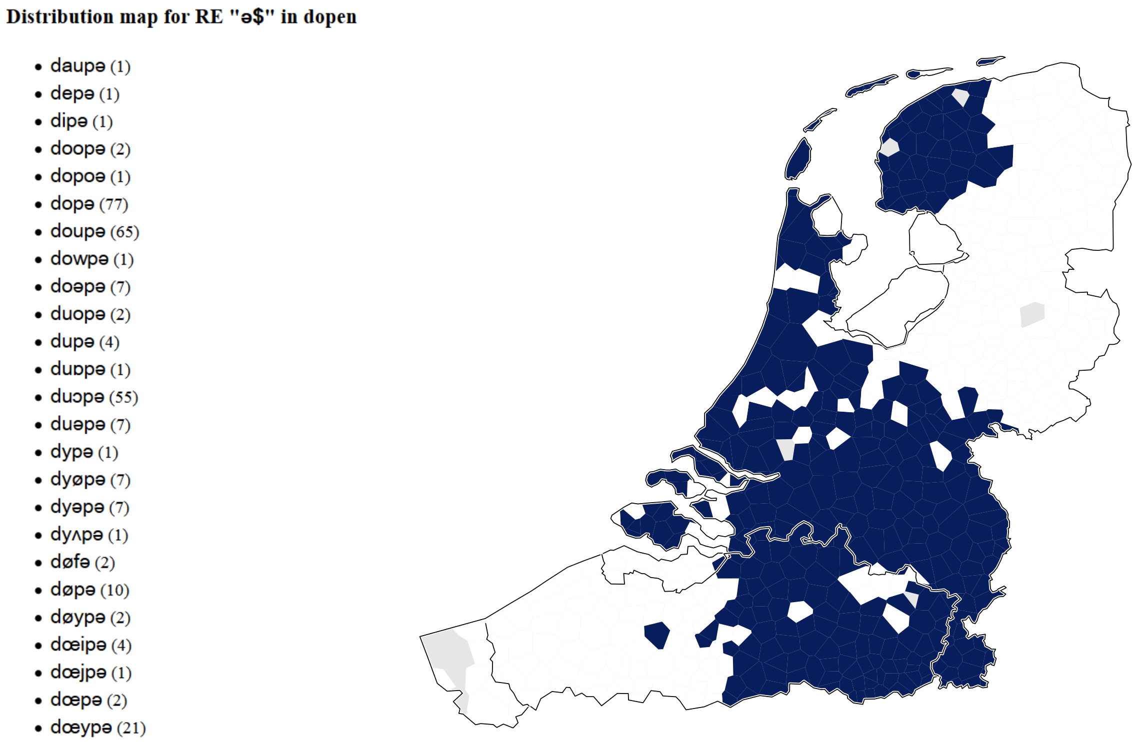

Figure 2. Map showing the distribution of pronunciations of the word *dopen* ('to baptize') ending in a schwa. To the left a part of the pronunciations selected by the regular expression is shown.

relative frequency of the character in the data at the given site. A map like this only gives a rough picture of the distribution of a speech sound, since the result depends on how well each data point has been sampled.[3] Still, the map can give a rough overview of the distribution of a dialect feature and/or of the quality of the data. It is striking that the chosen phonetic symbol in Figure 1 is almost completely lacking in the data from Belgium. When a pattern like this is found, it could either mean that the distribution of the specific feature very closely follows the national border, or, it could mean that it was not transcribed with the same phonetic symbol by transcribers of the Flemish and Netherlandic Dutch data. In fact this is one of the indications that the Dutch and Flemish fieldworker-transcribers did not use the phonetic alphabet (Wieling, Heeringa, and Nerbonne 2007) in the same way; it turned out that the Flemish fieldworker-transcribers used many fewer symbols. See Wieling and Nerbonne (2011b) for a suggestion on how to correct for the differences in phonetic alphabet using dialectometric techniques.

Distribution maps of specific words can also be created in Gabmap. By first choosing a variable (word) and then a specific variant (pronunciation) a map is created which shows where the chosen variant can be found. Regular expressions can also be used to create distribution maps. Figure 2 was created by first choosing the word *dopen* ('to baptize') and subsequently using the regular expression 'ə$' ('$' to mark end-of-string) for selecting all

[3]Sites with a lot of missing data could by coincidence get too high or too low relative frequencies compared to other sites.

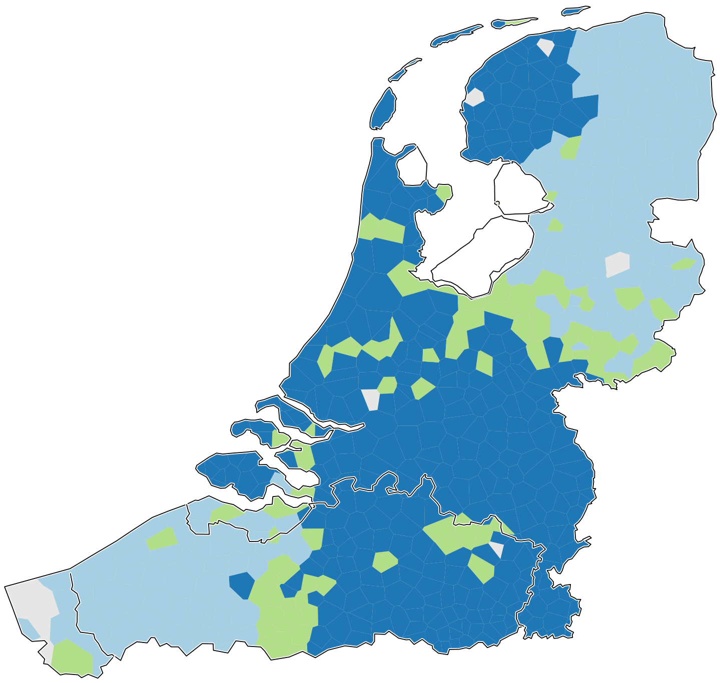

Figure 3. Map showing the distribution of three different types of endings for the word *dopen* in Dutch dialects: -*m* (light blue), -*n* (green) and *vowel* (dark blue). Gray spots are sites with missing data.

pronunciations ending with a schwa, illustrating one result of the weakening of unstressed syllables. In addition to creating the map, Gabmap shows a list of the chosen pronunciations.

The distribution maps in Gabmap can only show the presence or absence of a chosen feature. In traditional dialect maps, however, it is common to show the distribution of several different variants by using different symbols, patterns, or colors. For example, one might want to make a map of the word *dopen* showing the distribution of three different types of endings -*m* (e.g. [dopm]), -*n* (e.g. [dopən]) and ending in a *vowel* (e.g. [dopə]). This can be achieved in Gabmap by using a data file with a single variable (i.e. one data column):

	ending
Aalsmeer	vowel
Baardegem	-n
Coevorden	-m

When uploading the data, *categorical data* is used as data type and *binary comparison* as processing type. The map can be created as a *cluster map* in Gabmap. Since the clusters are coded in the uploaded data file, it does not matter which clustering algorithm is used, but the number of clusters should simply be the same as the number of different codes in the data file, which is 3 in the example case. The map is shown in Figure 3.

Aalsmeer NH — Aalten Gl

r	e	i	ɣ	ə		
r	e		x	ə	n	
		1	1		1	**3**

Figure 4. Example of computing of string edit distance.

2.3 Measuring linguistic distances

Dialectometric analyses are typically based on linguistic distances between pairs of sites in the data. The linguistic distances between sites are in turn calculated as the mean distances of the variables instantiated at both sites. Gabmap calculates these distances when a project is created. The distance measure used for string data is the *string edit distance* (or Levenshtein distance, Levenshtein 1966).

The string edit distance computes the minimal number of insertions, deletions and substitutions needed to change one character string into another. Gabmap computes the distance for all words and all pairs of sites and shows the alignments made (under *Measuring technique > alignments*). Figure 4 shows the alignment of the word *regen* ('rain') in the Aalsmeer dialect and the Aalten dialect. One deletion [i], one substitution [ɣ]∼[x] and one insertion [n] is needed for the alignment, which results in a distance of 3. The linguistic distance between two sites is the average of the distances of the words available from both sites.[4]

For other types of data other distance measures can be chosen. For numeric dialect data the Euclidean distance is used, and for categorical data either binary comparison or the 'Relative Identity Value' (*Gewichteter Identitätswert*, Goebl 2006, p. 416), a weighted similarity index, can be used. Instead of uploading actual dialect data it is also possible to upload a matrix of any kind of distances into Gabmap.

The distances are displayed in Gabmap as beam maps or network maps (see Nerbonne et al. 2011, p. 79). Another possibility is to display the distances from one site to all other sites (*references point maps*), which shows how similar or different the dialects might sound to a speaker of a specific dialect. Figure 5 shows a reference point map where Coevorden in the north-east of the Netherlands is the reference point.

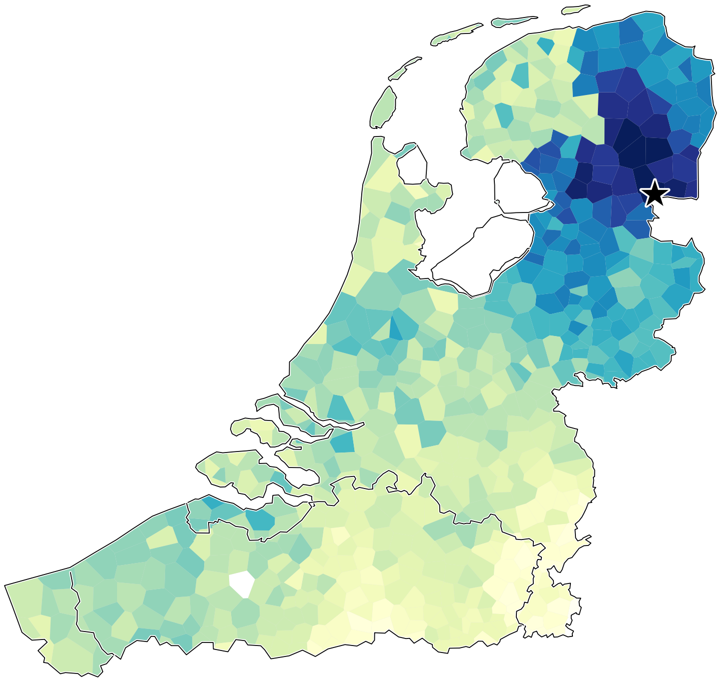

Figure 5. Reference point map. The lighter the color, the greater the linguistic distances from the starred reference site (Coevorden).

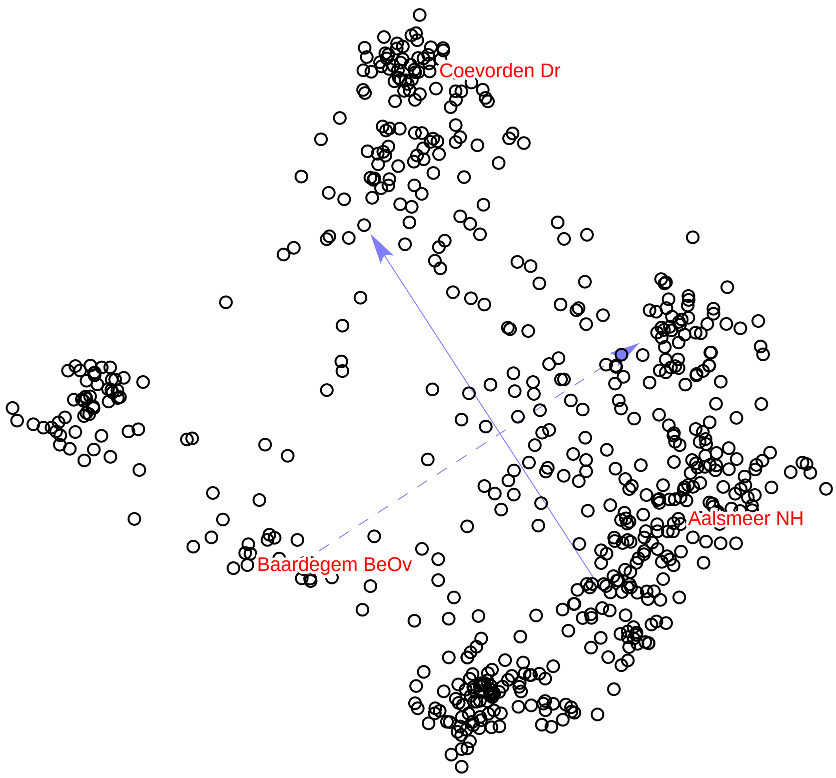

Figure 6. Plot of the result of multidimensional scaling in two dimensions. The labels of three reference sites are displayed.

2.4 Dialect continuum

Maps such as the reference point map in figure 5 only visualize the linguistic distances from one site to all other sites. In this map, there are very light areas to the north-east from the reference site Coevorden, as well as in the south of the language area. The map does not tell us whether these two areas are similar to or different from each other or not, only that both of them are linguistically very different from Coevorden. For an objective observer, a map that displays the linguistic relationships across all sites simultaneously might be more useful. This can be achieved by using multidimensional scaling (MDS).

MDS takes the full *sites* × *sites* distance matrix as input and creates a representation in an n-dimensional space where the distances are approximations of the original linguistic distances.[5]

This can be compared to trying to create a map using only the distances between cities in kilometers as information. The results of MDS can be plotted in a Cartesian coordinate system (*mds plots* in Gabmap). Similar data points will be close to each other in the plot.

An example of this is found in Figure 6, where the labels of three example sites have been added. The first dimension of an MDS analysis always explains as much as possible of the variance in the data, and additional dimensions add maximally to the precision of the approximation of the distances, but each additional dimension explains less of the variance than the previous one. In Figure 6, the solid arrow represents the first dimension explaining 49% of the variance in the data (correlation between the original linguistic distances and the Euclidean distances between the MDS coordinates: $r = 0.70$) and the dashed arrow represents the second dimension explaining 23% of the variance ($r = 0.48$).[6] Aalsmeer has the lowest value in the first dimension, while Baardegem has an intermediate value and Coevorden has a very high value. In the second dimension, on the other hand, Baardegem has a very low value, while Aalsmeer and Coevorden both have relatively high values. This means, that there are some linguistic features that Aalsmeer and Coevorden share (second dimension), but other features that are very different in these two dialects (first dimension). The plot clearly shows that there are some groups of dialects that cloud together, but also single sites which lie between those groups.

The results are easier to interpret if they are displayed on maps. The two first maps in Figure 7 show exactly the same results as Figure 6, but instead of displaying a coordinate

[4] If more than one pronunciation is available for a word from one site or both sites, an averaging procedure (ignoring identical pairs) is used (see Nerbonne and Kleiweg 2003, Sec. 3.2).

[5] On the use of multidimensional scaling in dialectology, see e.g. Embleton (1993), Heeringa (2004, pp. 156–161), Nerbonne (2011, pp. 487–489), and Embleton, Uritescu, and Wheeler (2013).

[6] If a map file is provided, the MDS plots produced by Gabmap are rotated using the Procrustes transformation (see, e.g., Peres-Neto and Jackson 2001), which has the effect that the sites presented in the MDS plot align with their geographic coordinates as closely as possible. The axes corresponding to the first two MDS dimensions are drawn on the graph.

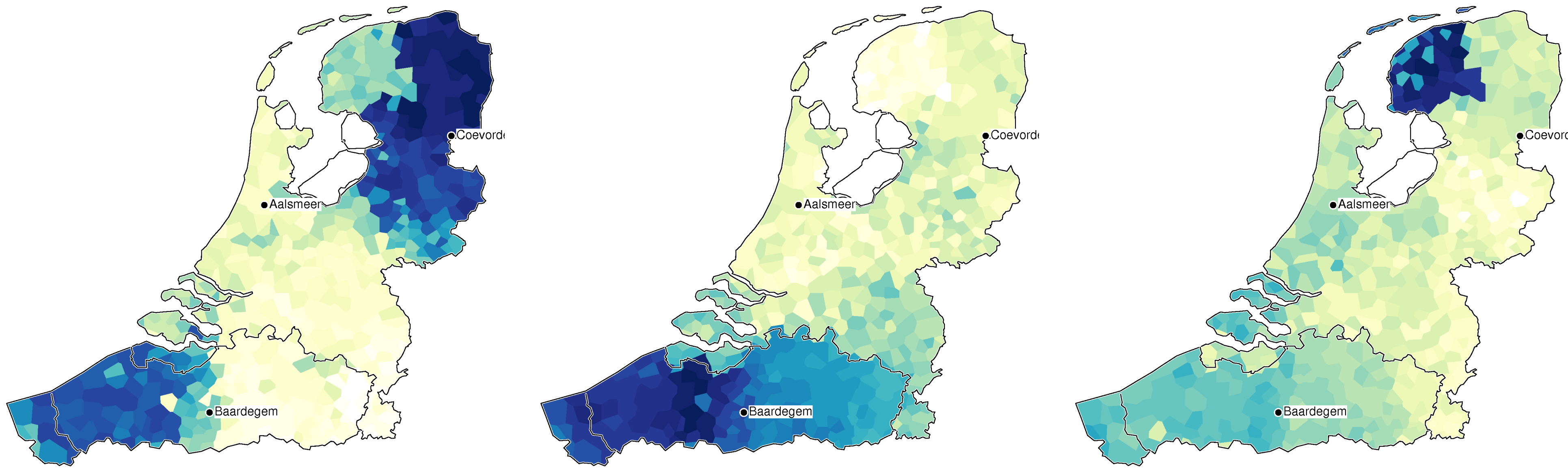

Figure 7. Maps of the first (left), second (center), and third (right) dimension of multidimensional scaling.

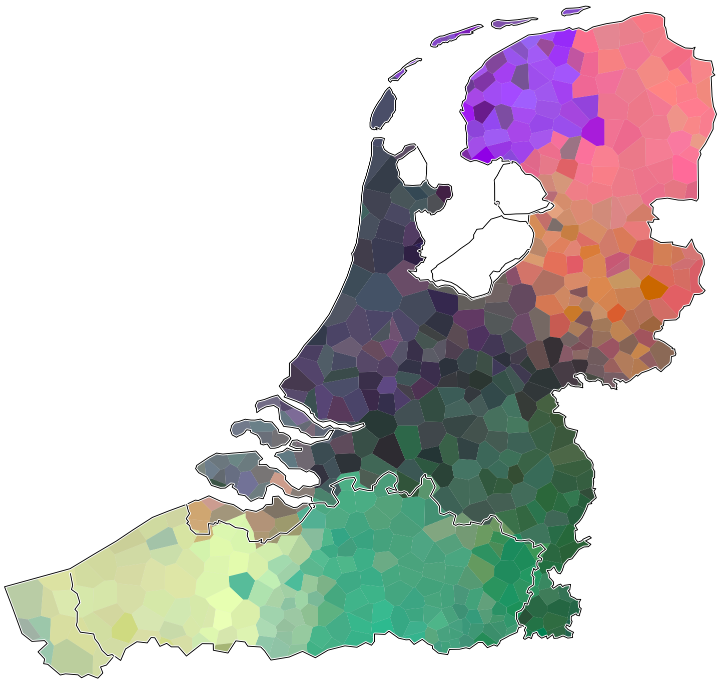

Figure 8. Map of multidimensional scaling applied to Dutch dialects ($r = 0.89$).

system, the area surrounding each site on the map has been colored according to the value of one dimension in the MDS analysis. The third map shows the third dimension ($r = 0.37$). Light color means high value, dark color low value. The maps show that the dimensions of the MDS represent different geographic distribution patterns: the first dimension shows a center–periphery effect, while the second dimension shows a northeast–southwest distribution. The third dimension mainly distinguishes Frisian (dark area) from the Dutch dialects. Multidimensional scaling to three dimensions has almost always explained around $80-90\%$ of the variance in the dialect data sets we have analysed, and it has been our experience that adding more than three dimensions to the analysis generally does not improve the solution much.

The maps in Figure 7 can be superimposed — or "put on top of each other" — using the red, green and blue (RGB) colors in order to show the aggregated dialectal differences, which gives the map in Figure 8.[7] All the maps of MDS results are found in *mds maps* in Gabmap. Similar colors in these maps indicate that the dialects share many features. The sharpest dialect border in Figure 8 is found in the north where the Frisian dialects are very different from the Dutch dialects. Frisian is in fact officially recognized as a separate, but closely related, language with its own written standard. The rest of the map shows less crisp borders, reflecting instead rather continuous transitions from one dialect area into the other.

[7]For a detailed account of how this is achieved, see Heeringa 2004, pp. 161–163, Leinonen 2010, pp. 207–208, and Nerbonne 2011, pp. 489–491.

2.5 Identifying dialect groups

The MDS plot in Figure 6 shows that despite the continuous nature of the dialect data, the dialects also seem to cluster together to some extent forming dialect groups. Dialectologists often want to be able to identify these kinds of dialect groups and draw borders between dialect areas on maps. We can seek groups of sites and dialect areas using cluster analysis. Clustering algorithms aim at minimizing the differences within each group of data points, while maximizing the distances across groups. Several so-called hierarchical clustering methods are available in Gabmap. Cluster analysis is applied to the distance matrix consisting of the pair-wise aggregate linguistic distances between places, and groups are formed based on similarity.[8]

The results of cluster analysis are shown in maps in Gabmap, where each cluster is displayed by a unique color.[9] Figure 9 shows the results of two different cluster algorithms: weighted average (left) and Ward's method (right). The contrast in these maps highlights the fact that different clustering algorithms have different biases and can lead to very different results. Ward's Method has a bias to favor equal size clusters, while weighted average is more faithful to the original linguistic distances. The figure shows that the map based on Ward's method has seven quite large clusters of dialects, while the map of weighted average has five large clusters and two very small ones.

Not ony do different clustering algorithms yield different results, each algorithm is also relatively unstable, meaning that small changes in the input data can lead to large changes in the cluster division. This is because each site is forced into a single cluster even in cases where the data might in fact be continuous. This can be compared to multidimensional scaling, which can show group structure in data, but also allows data points to float between groups or even show a truly continuous distribution (cf. Figure 6).

Because cluster analysis is a relatively instable method, *noisy clustering* (Nerbonne et al. 2008) has been implemented in Gabmap. In noisy clustering, cluster analysis is performed several times with different clustering methods and by contaminating the original distance matrix with (different) small amounts of random noise. The results of noisy clustering are displayed in a probabilistic dendrogram where percentages show how many times each cluster was encountered in the repeated clustering with noise. Clusters that appear in many runs of the analysis with added noise are particularly stable ones. For an example of noisy clustering in Gabmap, see Nerbonne et al. (2011, p. 83).

[8] For an introduction to cluster analysis and descriptions of differences between different cluster algorithms, see e.g. Jain and Dubes (1988), Manning and Schütze (1999, pp. 495–528), Heeringa (2004, pp. 146–156), and Prokić (2010, pp. 25–29).

[9] In contrast to MDS maps, the colors are arbitrary in the sense that similarity of colors does not indicate linguistic similarity. E.g. the light blue dialects are not necessarily any more similar to the dark blue dialects than to the red dialects.

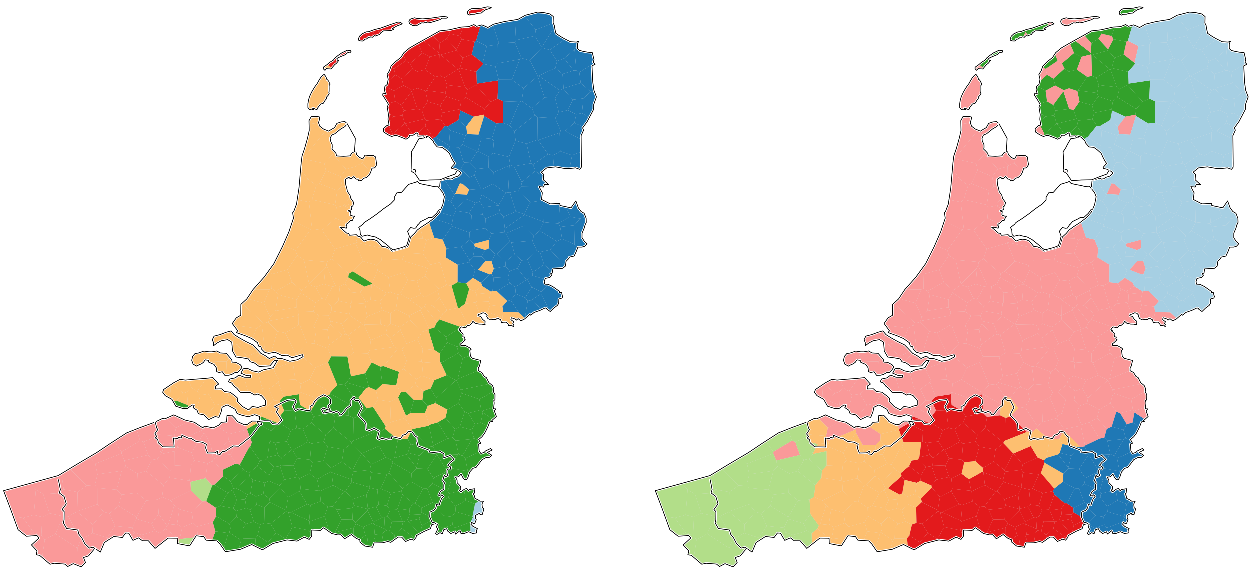

Figure 9. The results of two different cluster analyses: weighted average (left) and Ward's method (right). Seven clusters are displayed with seven distinct colors in both maps.

Another way of evaluating the results of cluster analysis is to compare the results of clustering to MDS (*cluster validation* in Gabmap). Figure 10 shows the two-dimensional MDS plot (Figure 6) colored according to the two different cluster analyses (Figure 9), respectively. Ward's method recognizes seven relatively large clusters, but at the cost of separating groups that are actually relatively similar according to the MDS (see, for example, the cloud of sites at the left side of the plot which belongs to one cluster according to weighted average but two different ones according to Ward's method). The two methods also disagree on how sites that fall between the clear clouds of sites are treated. Many of these are actually extreme points within a cluster, as indicated by the numbers added to the plots.

The comparison to MDS shows that, in this particular data set, the clusters might in fact not be as well separated on linguistic grounds as the cluster map might seem to suggest. Of course, the MDS plot only shows the first two dimensions of MDS which explain around 72% of the variance, so some of the information used in the cluster analysis is not accounted for in the MDS solution. For example, the third dimension of MDS singles out Friesland (cf. Figure 7). which will make it a more distinct cluster than the two first dimensions of MDS suggest. Hence, the amount of variance explained by different dimensions of MDS should also be considered when using MDS for validating cluster analysis.

2.6 Finding typical features or "shibboleths"

The dialectometric methods we discussed so far aim to find and characterize dialect groups at an aggregate level. A large number of variables (e.g. words) are used for investigating

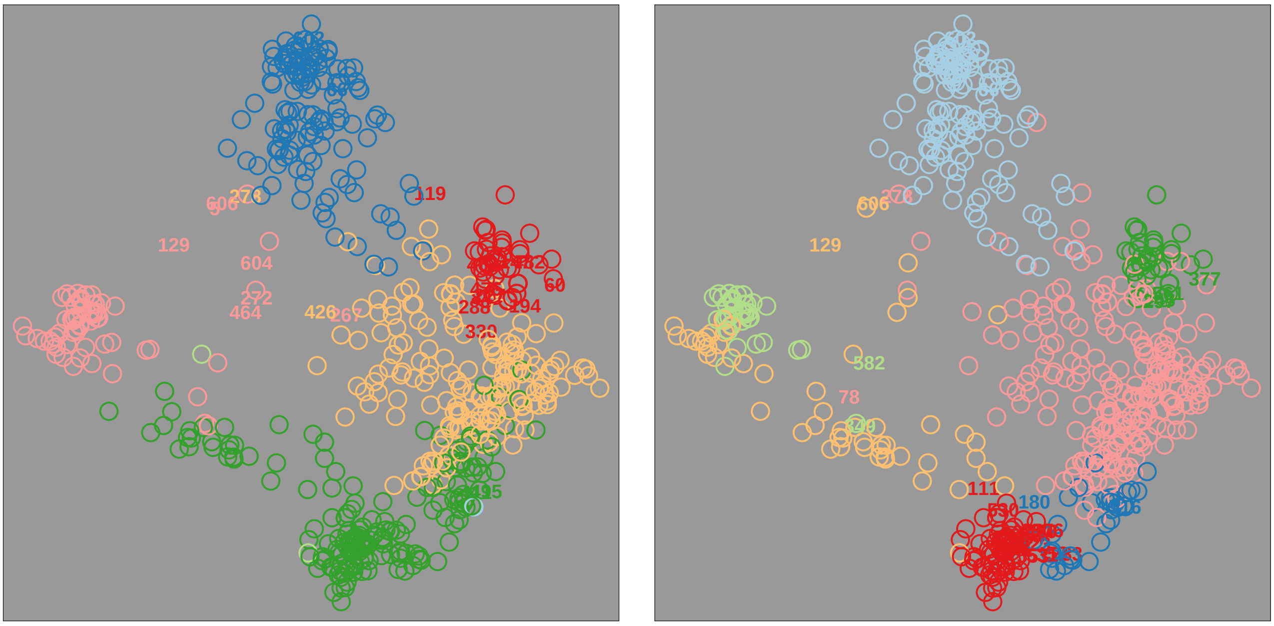

Figure 10. The results of weighted average (left) and Ward's method (right) compared to multidimensional scaling. The colors above correspond to those in Figure 9 (left and right, respectively).

overall dialectal differences. Often, we want to know which variable or variables are most characteristic for a specific dialect area. Such variables, termed shibboleths, referring to a variant of speech that betrays where a speaker is from (*Judges* 12:6), can be identified with the 'cluster determinants' function of Gabmap.

The cluster determinants option in Gabmap implements the method described in Prokić, Coltekin, and Nerbonne (2012).[10] The aim of the cluster determinants function is to find the items that are characteristic for a particular cluster, i.e. a set of sites. The method is related to the *Fisher's linear discriminant* (Schalkoff 1992, 90ff) and the information retrieval measures *precision* and *recall*. In essence, we would like to find items that distinguish sites in the target cluster from the sites outside it (possibly belonging to multiple clusters), but we also prefer the items that exhibit little variation within the target cluster. These two properties, *distinctiveness* and *representativeness*, together define how characteristic a particular item is for the target cluster.

Gabmap enables the investigation of typical linguistic elements ("cluster determinants") in three steps. In a first step, the target cluster is determined. The user can obtain a clustering using any of the clustering options described in Section 2.5, selecting one of the clusters as the target cluster. Even if more than two clusters are determined by this process, the important distinction is between the (selected) target cluster and the rest of the sites. The structure outside the target cluster is not used. Alternatively, the sites in the target cluster can be defined manually, e.g. based on theoretical motivations. The procedure also allows automatic clustering at first step, and adjusting the result manually.

[10]An earlier method based loosely on Wieling and Nerbonne (2011a) is also available for categorical data.

Table 1. The top- and bottom-three ranked 'shibboleths' for the Frisian cluster. The scores in the column *between* represents the differences between the Frisian cluster and the rest of the Dutch speaking area in our data set with respect to each item. The higher the score, the more *distinctive* the item. The scores in the column *within* measures the variation of the item within the Frisian area. The lower the score (variation), the more *representative* the item. The overall score at the column labeled *score* is the difference *between* − *within*.

Item	**between**	**within**	**score**
vinden	0.03	-2.37	2.41
knieën	1.13	-1.20	2.34
zoet	1.17	-1.12	2.29
nog	0.22	0.28	-0.06
kaf	0.63	0.72	-0.09
elf	0.27	0.36	-0.09

In a second step, the user selects the target cluster, and generates a ranked list of items along with their representativeness and distinctiveness scores. The scores are presented after normalization, so that the average (randomly selected) item would get a score of zero. The items are ranked based on an (equally weighted) linear combination of the two scores (see Prokić, Coltekin, and Nerbonne 2012, for the details of normalization and combination of the scores). Depending on the application, one may prefer to select the items based on just representativeness or just distinctiveness, or possibly on a differently weighted combination of the two. Gabmap allows downloading the resulting table, which the user may then experiment further with.

Table 1 presents the top three and bottom three shibboleth candidates for the Frisian area we discussed in the previous sections. The first item 'vinden' scores high because of the fact that it is pronounced uniformly within the Frisian area (low within score). However, it is definitely not a distinctive item (between score close to zero). The pronunciation differences between the Frisian cluster and the rest of the Dutch speaking area is almost exactly what would be expected from the differences measured in the whole data set with respect this item, i.e. quite similar pronunciations are found in other areas even though the exact same pronunciation does not occur outside Friesland. The other two top items show more balanced representativeness and distinctiveness scores. The least likely candidates all show small scores of distinctiveness or representativeness, and their combination result in low scores (around zero). An expert would already get a sense of specific items for a particular cluster by eyeballing the ranked list. However, the next step in 'Cluster determinants' function allows closer inspection of any item in the list.

In the last step, after determining the characteristic items, the user can select a particular item and visualize the differences with respect to this particular item using beam maps (see Nerbonne et al. 2011, p. 79), and list all forms (pronunciations) observed within or outside the target area along with their frequencies. For example, looking at the item *vinden* we identified as being representative of Frisian area, we observe that this item is pronounced as [finə] in all 52 sites in this area. The exact pronunciation is not found elsewhere in our larger area of interest. However, the distinctiveness score indicates that the pronunciation differences (as measured by Levenshtein distance) between Frisian area and the rest of the sites do not differ substantially. If we look at the second item in the list, *knieën*, we observe that the item varies within the Frisian cluster, in total we observe 15 forms of the item, and all except one of these forms are used exclusively within this cluster. The distinctiveness score also indicates that the pronunciation difference between Frisian area and the rest is over one standard deviation away from the typical pronunciation difference between two sites with respect to this item. Further inspection of the forms recorded within the Frisian area indicates that the pronunciation of *knieën* in this area almost always ends with an [s]. Similarly, all pronunciations of *zoet* (the third item in the list) in the Frisian area has a initial [s], while this is rare in other sites in our data.

3 User experiences

3.1 Some statistics

It is difficult to characterize the users of Gabmap in detail, as we decided against requiring users to identify themselves when developers of similar projects reported that mandatory registration appears to depress the enthusiasm for web applications. We can report that there were 45 users and 352 projects (excluding 10 guest users) as of late March, 2014. This figure ignores those with completed projects whose accounts expire after two months of no use (with one week's warning). The web server access for the last month indicates on average 2795 hits and 71 visits per day.

We have also presented tutorials on Gabmap at the *Nordic Congress of Dialectologists*, Uppsala, Aug. 20, 2010; at the *Tagung des Forums Sprachvariation*, Erlangen, Oct. 15, 2010; at the University of Potsdam, Dec. 7, 2010; in a poster at the 6th *International Conference on Language Variation in Europe* (ICLaVE), Freiburg, June 30, 2011; at *Digital Humanities* 2011 (Stanford) with about 12 participants; at the conference *Methods in Dialectology XIV* (London, Ontario, Aug. 2011) with 40 attending; at the conference *Comparing approaches to measuring linguistic differences* at the University of Gothenburg, Oct. 26, 2011; at the Society of Swedish Literature in Finland, Nov. 23-25, 2011; at the LOT winter school of the Dutch National Research School for Linguistics (Tilburg, Jan, 2012);

at a *Digital Humanities* summer school in Leuven, Sept., 2012 with roughly 10 participants; and at *Methods in Dialectology XV* conference (Groningen, Aug. 2014), with over twenty participants. Users have been pleased at the ease with which analyses can be conducted.

3.2 Examples of user work

Gabmap has been used for various purposes in the three years since it was first launched; these include not only linguistic and other research, but also the presentation of research to professionals and to interested popular science audiences.[11] The recent *Methods in Dialectology XV* conference included several talks which used Gabmap (and which are discussed below) as well as talks which compared treatments to Gabmap (e.g., talks by Simon Pickl and Fruzsina Vargha).

A number of users have especially exploited Gabmap's map-making facilities. Bouma and Hermans (2012) use Gabmap to project the distribution of different syllable onsets in medieval Dutch, and Wieling, Upton, and Thompson (2014) and Wieling (2013) use Gabmap's facilities for analyzing numerical data (lexical frequency differences) to provide analyses of the very large-scale BBC voice project. The work may be viewed in more detail at http://www.gabmap.nl/voices/ where users are encouraged to explore the lexical choices of all the respondents, or to contrast men's and women's speech or the speech of the young and old. Leinonen (in press) uses Gabmap's map-making facilites for analyzing data from the dictionary of Swedish dialects in Finland. She uses the clustering facility for making isogloss maps of single features with multiple variants as well as aggregating dialectometric maps.

Castro (2011), on the other hand, uses Gabmap's clustering routines in his argument that Southern Sui should be recognized as a separate dialect, distinct from Sandong Sui. Coloma (2012) focuses on just ten features in modern Spanish and, like Castro, exploits Gabmap's ability to process numeric data (differences in frequencies) and to invoke clustering and MDS. Scherrer (2012) introduces his own idea for measuring varietal distance based on comparing the number of identical lexicalizations in Swiss German dialect corpora to the number of cognates found there, and he uses Gabmap for MDS, clustering, and mapping even while examining the Cronbach's α score used in Gabmap to determine whether samples are large enough to provide a geographical signal and using a Mantel test comparing distance matrices determined using different techniques. Moran and Prokić (2013) investigated several endangered Dogon languages (spoken in Mali) emphasizing the need to preserve what is possible in communities with few speakers. They made use of Gabmap's

[11]Our thanks, too, to Erik R. Thomas, North Carolina, and Yonatan Belinkov, Tel Aviv, who referred us to their as yet unpublished work using Gabmap on Midwestern US varieties of English and on translations of the Hebrew Passover Haggadah, respectively.

probabilistic clustering routines as well as the mapping facilities. Reber (2013) focused not on dialect speech, but rather on the range of place names found at different settlements, i.e. the names of neighborhoods, fields, streets, paths, hills, peaks, rivers and other bodies of water. The author uses Gabmap for clustering and mapping.

Uiboaed et al. (2013) investigated corpus-based morphosyntactic dialectometry by first extracting corpus frequencies of various verbal “collostructions” (Stefanowitsch and Gries 2003) in Estonian and then examining the results for geographic cohesion using both correspondence analysis and Gabmap’s clustering routines.

Mathussek (2013, pp. 248-251) uses Gabmap’s aggregating, dialectometric focus to analyze middle Franconia (in northwest Bavaria) and to contrast the aggregate views with perspectives from traditional research and from perceptual dialectology. The dialectometric approach was crucial in identifying field worker boundaries in the data, which led her to ignore phonetic details (diacritics) before proceeding, an issue which is the focus of Mathussek (submitted). Mathussek’s approach is emphatically pluralistic, and she notes that it was the failure of initial dialectometric analyses to agree with traditional ones that led her to pursue the possibility of field worker confounds. Mathussek (2014, Chap.4) discusses Gabmap as means of returning to older data sets with new techniques — naturally, in order to obtain new insights, or at least to examine older ideas from a fresh vantage point.

Šimičić et al. (2013) analyzed coastal Croatian dialects but also varieties from the Italian provinces of Molise, attending to phonological and lexical variation. The two linguistic levels correlated strongly ($r = 0.72$), and the authors interpret the differences to be due to the stronger historical signal in phonology, and the greater volatility of the lexicon. Due to the complicated history of Croatian migrations, one might have expected the usual dialect areas and dialect continua not to emerge, and they indeed do not emerge from this analysis. Instead the analysis uncovers a great many discontinuities, particularly on the northern island of Istria, which the authors suggest ought to be attributed to migration. The Štokavian and Čakavian varieties of the south were less diverse, and the varieties spoken in Molise, Italy were very distinct from the others. The authors conclude methodologically that the aggregating view inherent in Gabmap has advantages over the traditional analyses based on isoglosses, in particular because it obviates the need to choose which isoglosses are to be regarded as probative.

Mitterhofer (2013) used Gabmap to identify cognates and other related words in varieties of Bena in Tanzania, comparing Gabmap’s edit-distance measures to the Summer Institute of Linguistics’ “Survey on a Shoestring” (1990), and Bloem et al. (submitted) uses the "cluster determinants" feature of Gabmap to identify characteristic mispronunciations in foreign accents in English.

Snoek (2013) uses Gabmap to research lexical relations among Athapaskan languages in order to improve the understanding of their historical relations, and Snoek (2014) provides an article-length review of Gabmap targeted at researchers in language documentation. The author analyzes phoneme strings denoting body-part terms in Northern Athapaskan languages (in Canada and Alaska). The application of dialectometrical tools is appropriate for these Athapaskan languages because their relations to one another are poorly established in Amerindian scholarship. He adds to existing documentation by explaining how maps may be produced for Gabmap using Google Earth, and he has some important warnings about how Gabmap may handle transcriptions involving digraphs or trigraphs. Most intriguingly, he shows how Gabmap's data examination facilities may be very useful even when researchers do not aim at a quantitative analysis of their data. He concludes that "Gabmap is excellent software that permits the mapping and comparison of linguistic data in a fast and generally painless manner."

4 Conclusions

Gabmap offers a range of processing possibilities all geared to highlighting and tallying linguistic differences. Nerbonne et al. (2011) sketched some of these, and the current paper aims to supplement that one by describing other possibilities and also to review some of the uses to which Gabmap has been put.

Gabmap would undoubtedly benefit from further use and also from the incorporation of various advances in dialectometry since 2010, including more sensitive measures for pronunciation differences that incorporate segment differences (Wieling, Margaretha, and Nerbonne 2012; List 2012), and we have noted that tutorial material as well as reference material at various levels is invaluable. More material would be precious. The phylogenetic grouping procedures such as NeighborNet or Bayesian Monte Carlo Markov-Chain techniques (Felsenstein 2004, Ch.16,18,35) are valuable historical perspectives for many of the questions dialectologists entertain. The maps Gabmap produces are not geo-referenced, and this handicaps some interesting applications involving comparing the diffusion of linguistic culture with other sorts of culture (Manni et al. 2008).

Gabmap is also open-source, and we would welcome proposals from others to incorporate further processing possibilities into Gabmap, although we are also wary of the time that might be needed to see this through successfully.

References

Bloem, J. et al. (submitted). "Automatically Identifying Characteristic Features of Non-Native English Accents". In: *The Future of Diaelcts*. Ed. by M.-H. Côté, R. Knooihuizen, and J. Nerbonne. Language Variation 1. Berlin: Language Science Press.

Bouma, G. and B. Hermans (2012). “Syllabification of Middle Dutch”. In: *The Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2)*. Ed. by F. Mambrini, M. Passarotti, and C. Sporleder. Lisbon, pp. 27–38.

Castro, A. (2011). “Southern Sui: a Fourth Sui Dialect”. In: *Journal of the Southeast Asian Linguistics Society* 4.2, pp. 1–31.

Coloma, G. (2012). “The Importance of Ten Phonetic Characteristics to Define Dialect Areas in Spanish”. In: *Dialectologia: revista electrònica* 9, pp. 1–26.

Embleton, S. (1993). “Multidimensional Scaling as a Dialectometrical Technique. Outline of a Research Project”. In: *Contributions to Quantitative Linguistics*. Ed. by R. Köhler and B. Rieger. Language and Space. Dordrecht: Kluwer, pp. 267–276.

Embleton, S., D. Uritescu, and E. S. Wheeler (2013). “Defining Dialect Regions with Interpretations: Advancing the MDS Approach”. In: *LLC: The Journal of Digital Scholarship in the Humanities* 28.1, pp. 13–22.

Felsenstein, J. (2004). *Inferring Phylogenies*. 2nd ed. Sinauer Associates Sunderland.

Goebl, H. (2006). “Recent Advances in Salzburg Dialectometry”. In: *LLC: The Journal of Digital Scholarship in the Humanities* 21.4, pp. 411–435.

Goeman, T. and J. Taeldeman (1996). “Fonologie en morfologie van de Nederlandse dialecten: Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten”. In: *Taal en Tongval* 48, pp. 38–59.

Heeringa, W. (2004). “Measuring Dialect Pronunciation Differences using Levenshtein Distance”. PhD thesis. University of Groningen.

Jain, A. K. and R. C. Dubes (1988). *Algorithms for Clustering Data*. Upper Saddle River, New Jersey: Prentice-Hall, Inc.

Leinonen, T. (2010). “An Acoustic Analysis of Vowel Pronunciation in Swedish dialects”. PhD thesis. University of Groningen.

— (in press). “Dialektgeografi och dialektometri”. In: *Sociolingvistik i praktiken*. Ed. by S. Boyd and S. Ericsson. Lund: Studentlitteratur.

Levenshtein, V. I. (1966). “Binary Codes Capable of Correcting Deletions, Insertions, and Reversals”. In: *Soviet Physics Doklady* 10, pp. 707–710.

List, J.-M. (2012). “Multiple Sequence Alignment in Historical Linguistics”. In: *Proceedings of ConSOLE XIX*, pp. 241–260.

Manni, F. et al. (2008). “Do Surname Differences Mirror Dialect Variation?” In: *Human Biology* 80.1, pp. 41–64.

Manning, C. D. and H. Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT press.

Mathussek, A. (2013). “Sprachräume im Mittelfranken”. In: *Handbuch zum Sprachatlas von Mittelfranken. Dokumentaiton und Auswertung*. Ed. by H. H. Munske and A. Mathussek. Schriften zum Bayerischen Sprachatlas. Bnd 9. Heidelberg: Universitätsverlag Winter.

Mathussek, A. (2014). *Sprachräume und Sprachgrenzen im Untersuchungsgebiet des Sprachatlas von Mittelfranken: Traditionelle Dialektgeographie, Wahrnehmungsdialektologie, Dialektometrie*. Heidelberg: Universitätsverlag Winter.

— (submitted). "On the Problem of Field Worker Isoglosses". In: *The Future of Diaelcts*. Ed. by M.-H. Côté, R. Knooihuizen, and J. Nerbonne. Language Variation 1. Berlin: Language Science Press.

Mitterhofer, B. (2013). *Lessons from a Dialect Survey of Bena: Analyzing Word Lists*. SIL International Electronic Survey Report 2013-020. Summer Institute of Linguistics.

Moran, S. and J. Prokić (2013). "Investigating the Genealogical Relatedness of the Endangered Dogon Languages". In: *LLC: The Journal of Digital Humanities Scholarship* 28 (4), pp. 676–791.

Nerbonne, J. (2011). "Mapping Aggregate Variation". In: *Language Mapping*. Ed. by A. Lameli, R. Kehrein, and S. Rabanus. Language and Space 2. Berlin: De Gruyter, pp. 476–501.

Nerbonne, J. and P. Kleiweg (2003). "Lexical Distance in LAMSAS". In: *Computers and the Humanities* 37.3, pp. 339–357.

Nerbonne, J. et al. (2008). "Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering". In: *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society*. Ed. by C. Preisach et al. Berlin: Springer, pp. 647–654.

Nerbonne, J. et al. (2011). "Gabmap – A Web Application for Dialectology". In: *Dialectologia* Special Issue II, pp. 65–89.

Peres-Neto, P. R. and D. A. Jackson (2001). "How well do Multivariate Data Sets Match? The Advantages of a Procrustean Superimposition Approach over the Mantel Test". In: *Oecologia* 129.2, pp. 169–178.

Prokić, J., C. Coltekin, and J. Nerbonne (2012). "Detecting Shibboleths". In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, 72—80.

Prokić, J. (2010). "Family Resemblances". PhD thesis. University of Groningen.

Reber, J. (2013). "Strukturen und Muster in der Namenwelt: Quantitative und qualitative Untersuchungen zum Toponymenbestand der beiden Solothurner Amteien Dorneck-Thierstein und Olten-Gösgen". PhD thesis. Universität Basel.

Schalkoff, R. J. (1992). *Pattern Recognition*. New York: Wiley & Sons.

Scherrer, Y. (2012). "Recovering Dialect Geography from an Unaligned Comparable Corpus". In: *In proceedings of the EACL Workshop on Visualization of Language Patters and Uncovering Language History from Multilingual Resources (LINGVIS & UNCLH)*.

Šimičić, L. et al. (2013). "Diatopic Patterning of Croatian Varieties in the Adriatic Region". In: *Journal of Slavic Linguistics* 21.2, pp. 259–301.

Snoek, C. (2013). “Using Semantically Restricted Word-Lists to Investigate Relationships among Athapaskan Languages”. In: *Approaches to measuring linguistic differences*. Ed. by L. Borin and A. Saxena. Trends in Lingusitics Studies and Monographs 265. Germany: de Gruyter Mouton, pp. 231–248.

— (2014). “Review of Gabmap: Doing Dialect Analysis on the Web”. In: *Language Documentation and Preservation* 8, pp. 192–208.

Stefanowitsch, A. and S. T. Gries (2003). “Collostructions: Investigating the Interaction of Words and Constructions”. In: *International Journal of Corpus Linguistics* 8.2, pp. 209–243.

Uiboaed, K. et al. (2013). “Variation of Verbal Constructions in Estonian Dialects”. In: *Literary and Linguistic Computing* 28.1, pp. 42–62.

Wieling, M. (2013). “Voices Dialectometry at the University of Groningen”. In: *Analysing 21st-century British English: Conceptual and Methodological Aspects of the BBC ’Voices’ Project*. Ed. by C. Upton and B. Davies. London: Routledge, pp. 208–218.

Wieling, M., W. Heeringa, and J. Nerbonne (2007). “An Aggregate Analysis of Pronunciation in the Goeman-Taeldeman-van Reenen-Project Data”. In: *Taal en Tongval* 59.1, pp. 84–116.

Wieling, M., E. Margaretha, and J. Nerbonne (2012). “Inducing a Measure of Phonetic Similarity from Pronunciation Variation”. In: *Journal of Phonetics* 40.2, pp. 307–314.

Wieling, M. and J. Nerbonne (2011a). “Bipartite Spectral Graph Partitioning for Clustering Dialect Varieties and Detecting their Linguistic Features”. In: *Computer Speech & Language* 25.3, pp. 700–715.

— (2011b). “Measuring Linguistic Variation Commensurably”. In: *Dialectologia* Special Issue II. Spec. Iss. on Production, Perception and Attitude ed. by J. Nerbonne, S. Grondelaers, D. Speelman & M.-P. Perea, pp. 141–162.

Wieling, M., C. Upton, and A. Thompson (2014). “Analyzing the BBC Voices data: Contemporary English Dialect Areas and their Characteristic Lexical Variants”. In: *LLC: The Journal of Digital Scholarship in the Humanities* 29.1, pp. 107–117.