***Dialectologia.* Special issue, II** *(2011), 65-89.*
ISSN: 2013-2247

Received 7 January 2011.
Accepted 11 May 2011.

GABMAP — A WEB APPLICATION FOR DIALECTOLOGY[1]

John NERBONNE, Rinke COLEN, Charlotte GOOSKENS, Peter KLEIWEG,
and Therese LEINONEN
Center for Language and Cognition, University of Groningen
{j.nerbonne,c.s.gooskens,p.c.j.kleiweg}@rug.nl, rinke.colen@gmail.com,
therese.leinonen@gmail.com

Abstract

Gabmap [2] is a web application aimed especially to facilitate explorations in quantitative dialectology — or dialectometry — by enabling researchers in dialectology to conduct computer-supported explorations and calculations even if they have relatively little computational expertise. Gabmap creates various views of dialect data, from histograms of characters used to spot coding errors, to alignments of phonetic transcriptions used in measuring pronunciation distance, to colored multi-dimensional scaling plots intended to illustrate quantitative results insightfully. Many analyses are accompanied by facilities allowing researchers to probe further, e.g. seeking the most important linguistic bases of an areal division, or examining the results of clustering for statistical reliability. These are also intended to inform the critical discussion of quantitative techniques, i.e. a comparision between quantitative analyses and non-quantitative (qualitative) work. For this reason Gabmap also includes support for qualitative analyses, such as facilities to map the occurrence of individual features. The software is in use, and the source code is openly available.

[1] The development of Gabmap was supported in 2010 by grant CLARIN-NL-09-014 from the CLARIN-NL program (http://www.clarin.nl/) to the ADEPT project (Assaying Differences using Edit Distance of Pronunciation Transcriptions), which we acknowledge gratefully. CLARIN-NL participates in the European CLARIN program (http://www.clarin.eu), whose aim is to develop a general infrastructure for scientific applications of language and text processing.
[2] Gabmap is accessible at http://www.gabmap.nl/

©Universitat de Barcelona

Keywords
computational linguistics, dialectology, dialectometry, quantitative linguistics, web application, maps, edit distance

GABMAP — UNA APLICACIÓN WEB PARA LA DIALECTOLOGÍA

Resumen

Gabmap es una aplicación destinada especialmente a facilitar los trabajos en dialectología cuantitativa —o dialectometría— permitiendo a los dialectólogos, incluso a los que tienen pocos conocimientos de técnicas computacionales, llevar a cabo análisis y cálculos asistidos por ordenador. Gabmap ofrece visualizaciones diversas de los datos dialectales, desde histogramas de los caracteres utilizados para detectar errores de codificación, a alineaciones de las transcripciones fonéticas usadas en la medida de la distancia en la pronunciación, así como gráficos multidimensionales coloreados destinados a ilustrar cuantitativamente los resultados. Muchos análisis van acompañados de herramientas que facilitan nuevas investigaciones; por ejemplo, buscando las bases lingüísticas más importantes de una división de área, o examinando los resultados de los conglomerados a partir de su fiabilidad estadística. Estos también pretenden aportar una discusión crítica a las técnicas cuantitativas; por ejemplo; una comparación entre análisis cuantitativos y no cuantitativos (cualitativos). Por esta razón, Gabmap incluye soporte para análisis cualitativos y herramientas para cartografiar las ocurrencias de rasgos individuales. El software utilizado es de código abierto.

Palabras clave
lingüística computacional, dialectología, dialectometría, lingüística cuantitativa, aplicación de mapas en Web, mapas, edición de la distancia lingüística

1. Introduction and motivation

1.1. Scientific motivation

The study of linguistic variation — especially dialectal (geographical) variation, but also social variation of different sorts — has held a central position in linguistics for well over a century. The last two decades have witnessed enormous progress in the quantitative analysis, i.e., the automatic measurement of linguistic differences (DIALECTOMETRY), which yields reliable and valid characterizations, e.g. when a hundred or so words are sampled at a few dozen or more sites (Goebl 2006, Nerbonne 2009, Nerbonne & Heeringa 2010, Goebl 2010).

©Universitat de Barcelona

The fundamental motivation for dialectometry lies in the opportunity to AGGREGATE large amounts of dialectal data. As Goebl has put it, this "condenses" (*verdichtet*) the data, strengthening the signals of speaker provenance. It also offers an alternative approach — if not a complete answer — to the long-standing problems of the relations between isoglosses and dialect areas. As Bloomfield (1927: 328) notes "isoglosses rarely coincide along their whole extent". See Bloomfield for a discussion of Kloeke's (1927) book-length discussion on differences in the isoglosses associated with 'house' and 'mouse,' which were identical in early Germanic. Chambers and Trudgill (1998) discuss further examples, e.g. from French, but they conclude the lack of an account of this relation as a "notable weakness in dialect geography" (p. 97). The opportunity to aggregate substantial amounts of data also opens dialectology to the deployment of statistical analysis and to the use of representative samples. Further, providing computational facilities within which to experiment with quantitative and qualitative analyses contributes to the replicability of the analytical tools used in the discipline. Aurrekoetxea and Ormaetxea (2010) is a recent compilation of papers on dialectometric techniques and emerging research questions.

Although Gabmap attempts to provide useful facilities for dialectologists of different theoretical and methodological persuasion, Gabmap is particularly well suited for the analysis of phonetic transcriptions using string comparison algorithms, a type of analysis we have long championed (see Nerbonne & Heeringa 2010, and references there). Nerbonne et al. (2010) argue that analyses comparing phonetic transcriptions effectively compare each phonetic segment separately and automatically, which means that the resulting analyses are (i) more reliable because they are based on more data; (ii) easier to implement because they obviate the manual step of "appraisal" (Goebl's *Taxierung*) in which items of comparison are abstracted from data collections and categorized for later analysis (so that perhaps only the vowel is used from transcriptions such as [naɪt^{h}] or [nat], Eng. 'night'); and (iii) somewhat less biased than atlas materials analyzed at a categorical level because they involve the comparison of material that is essentially randomly chosen, namely all the segments in words that were not the primary motivation for inclusion in the dialect atlas's set of words.

©Universitat de Barcelona

1.2. Previous work

Dialectometry has not enjoyed wide use due to its demanding technical threshold, requiring special software installations, some of which have their own pre-requisites. The most popular package is Haimerl's MS Windows-based *Visual Dialectometry* (VDM), which has been used extensively in studying the dialectology of Romance languages (Haimerl 1998, 2006). We are impressed by this work and provide some of its facilities (notably "reference point maps"), but we attempted to supersede it both in dialectometrical range but also with respect to general facilities which should be of interest to dialectologists. RuG/L04[3] is a UNIX-flavored package developed by one of us (Peter Kleiweg) at the University of Groningen which differs from VDM in offering facilities for comparing transcriptions and in some mapping techniques. It runs on several platforms.

We hope that the general facilities will help make Gabmap useful to working dialectologists, including those who would prefer not to work dialectometrically.

1.3. Goals and intended users

Gabmap has been developed to make dialect analysis tools available to working dialectologists and other students of linguistic variation in an easy-to-use web application. In addition to dialectometric analyses, Gabmap generates various data summaries, supporting error detection in input data, providing researchers with useful overviews, and enabling the creation of distribution maps of any number of linguistic variables — words, morphological realizations, and also phonetic characters or patterns, depending on the user's data. In this respect Gabmap goes well beyond dialectometry, supporting the exploration of a large number of user-defined variables in different ways.

Gabmap allows linguists to upload their variationist data in different formats, but in particular, in the form of tab-separated values, which are easily provided from spreadsheets, which are popular systems for linguistic data collection and organization. Various overviews of the data are created automatically in order to support users who wish to explore freely. Tools are made available to support the creation of maps from Google Earth,™ to convert different character encodings in input data into Unicode IPA

[3] See http://www.let.rug.nl/kleiweg/L04/

©Universitat de Barcelona

(UTF-8 or UTF-16), the "native" format in Gabmap, and also to exploit selected statistical routines using R.[4]

The heart of Gabmap is the measurement of differences, which may be categorical (e.g., different lexical realizations of one concept or different forms of one affix), numerical (e.g., sets of formant frequencies for vowels), or string based (e.g., phonetic transcriptions). Although various options are supported, we attempted to identify sensible defaults for inexperienced users throughout. Differences in linguistic items are then aggregated to obtain a robust characterization of the relations among the sites (or other groups of speakers), and these are analyzed and projected onto various sorts of maps to support scholarly investigation. Figure 1 provides a sample of the sorts of analyses and cartographic projections Gabmap provides. Because traditional dialectology emphasized dialect areas, i.e. areas of relative linguistic uniformity as the most important organizing element in dialectology, particular attention is paid to techniques for identifying natural groups (of sites) in data, examining them critically, and extracting the most representative and distinctive variables in them.

Although it is not our focus in what follows, we add here that the routines which seek natural groupings and affinities among dialects do *not* assume that the groupings are geographically based. They might therefore just as well be applied to variationist data to investigate non-geographic conditioning, e.g. social, sexual or ethnic differences. So while our emphasis has clearly been the development of software to support dialect geography, it is straightforward in Gabmap to perform "dialectometric" analyses of other variationist data, e.g. to see whether aggregate pronunciation distances distinguish two social groups.

[4] See http://www.r-project.org/

©Universitat de Barcelona

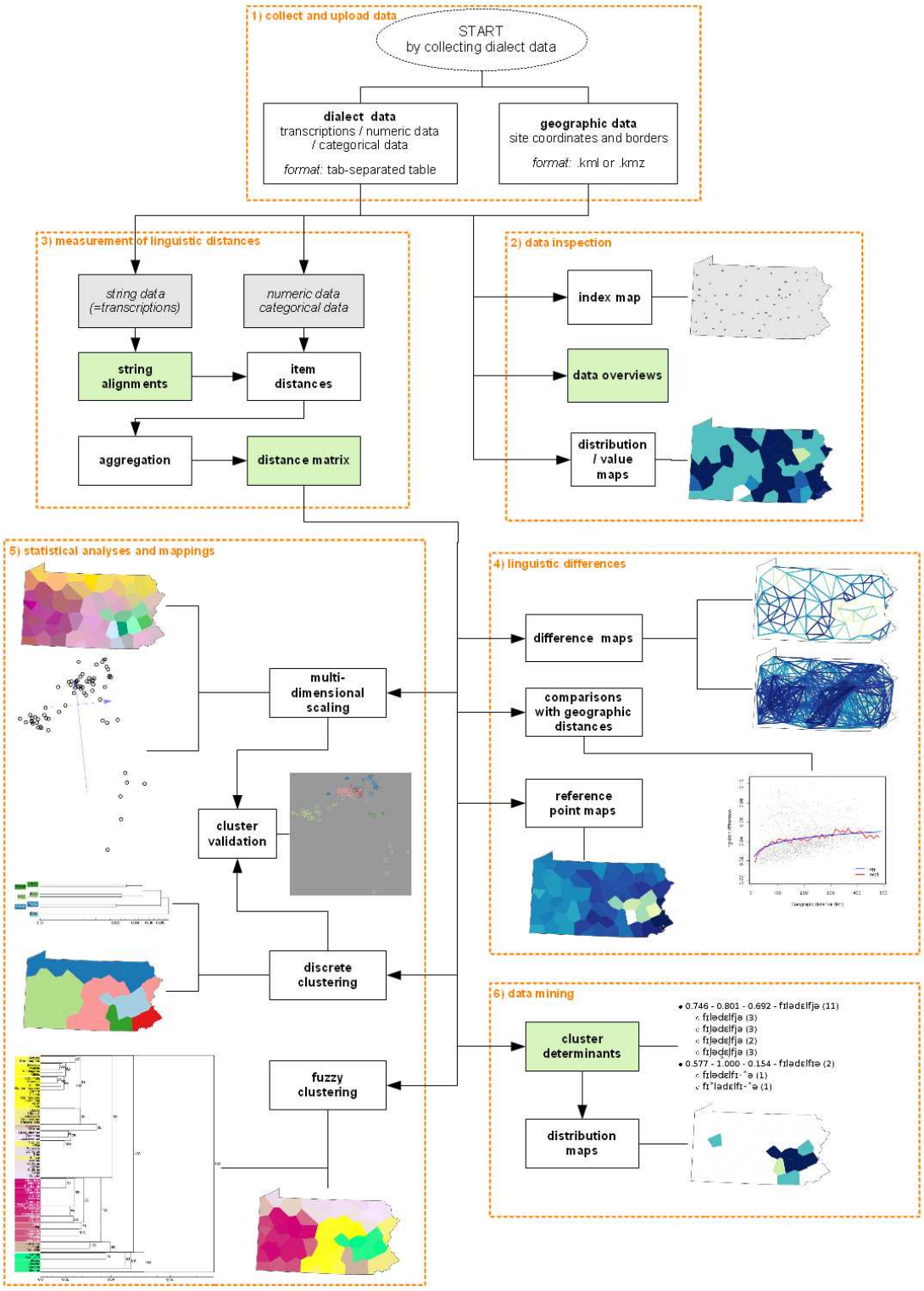

Figure 1. A sketch of some of the processing outputs supported by Gabmap. See Section 3 “Walk Through” for details.

©Universitat de Barcelona

2. Design choices and Implementation

In this section we present the technical realization of Gabmap, including its background, its input-output behavior, concerns about "piloting" inexperienced users, and its implementation.

2.1. Background

We implemented Gabmap after developing and supporting the RuG/L04 package for dialectometry since 2004.[5] L04 enjoyed limited use — primarily among students of dialectology, but to a limited extent, among others, e.g., population geneticists. From feedback from users, some of whom traveled to Groningen to learn to use the package, we knew that the UNIX command-line interface was found forbidding, and that the very large number of options supported confused most (potential) users. We set as goals for Gabmap therefore that its user interface be menu based and also that sensible defaults be chosen for as many analysis steps as possible. The goal was to allow users freedom to try alternatives, but at the same time to guide them toward sensible choices (e.g. in the sort of clustering techniques used).

The implemented web-interface enlarges the range of opportunities for dialectologists in several respects if we compare it to the RuG/L04-software, not only providing a more user-friendly interface, but in fact offering processing facilities for more complex tasks than can be carried out in RuG/L04. The web-interface is realized using a large number of scripts (see below) some of which directly implement new procedures (ones not in RuG/L04) in various programming languages and some of which invoke programs such as R in order to provide additional functionality. This illustrates the advantage of web-applications over traditional software distributions noted above, namely that developers control the configuration of the machine on which the software runs and need not assume specific configurations on user machines. The end result for us was a package that is much more than a new user interface, in fact, a new package that exploits RuG/L04 components where possible.

[5] See http://www.let.rug.nl/kleiweg/L04/

©Universitat de Barcelona

2.2 *Input/Output*

Input (linguistic) data is accepted in tabular form, e.g. via spreadsheets, which in our experience are popular (initial) data management systems used by researchers in dialectology. Gabmap may be used to analyze categorical data (lexical or syntactic data), or numerical data (vectors of formant frequencies of vowels), but is perhaps most interesting when applied to the analysis of phonetic transcriptions either in Unicode (UTF-8 or UTF-16) or X-SAMPA (a conversion tool is supplied to convert X-SAMPA to Unicode). In addition, linguistic differences (provided in tabular form) obtained from other analysis software may be further analyzed for geographical coherence and/or projected to maps. The interfaces are defined to allow the use of separate components wherever that seemed sensible.

Since the core topic of dialectology is the distribution of linguistic variation as influenced by geography, special attention is paid to the problem of obtaining and using maps. Instructions are provided for extracting maps from Google Earth,[6] and a program is made available which converts site names with longitude-latitude coordinates to .kml format.

Graphical output is provided in PostScript with conversions standardly available in PDF and PNG. We have taken pains to provide output appropriate for black-and-white printing wherever feasible, as researchers are still often unable to publish in color without incurring exorbitant additional costs. In addition to graphical output, Gabmap also provides output in tabular form.

2.3 *Interacting with users*

Since we aim to provide relatively sophisticated computational facilities *inter alia* to computationally inexperienced users, issues of how to interact with users arise frequently. On the one hand, we did not wish to proselytize, imposing our own scientific views on users. But, as noted above, our experience with users of RuG/L04 suggested that they were overwhelmed by the range of technical choices they might make, and we witnessed users in earlier trials who appeared to simply try everything until they found analyses they found congenial — users who appeared to “shop” about for analysis

[6] http://www.google.com/earth/

©Universitat de Barcelona

techniques. We were therefore concerned that we encourage users to focus on reliable techniques and that we discourage their shopping among techniques until they obtained results of sort they wished to see. Researchers who are new to computational analysis are often unfamiliar with the "embarrassment of riches" the techniques provide — the many minor variants of techniques that are often easily implemented and which yield scientifically rather different results.

Two processing steps may be adduced as illustrations of the dangers of the embarrassment of (analytical) riches: clustering on the one hand and variants of string alignment and string distance on the other. There is great interest in applying clustering to dialect data, as traditional results normally divide sites into groups, or DIALECT AREAS, meaning that clustering facilitates the comparison to older findings. But as is well known, clustering — seeking groups in data — is unstable, meaning that small differences in input data may lead to large differences in results (Kleinberg 2003; Prokić and Nerbonne 2008). There are, moreover, dozens of clustering algorithms, often yielding very different results, making it scientifically unsatisfactory for a researcher to simply check a number of results for one that he finds appealing. In Gabmap we have included clustering validation facility that allows users to compare clustering results to a plot obtained via multi-dimensional scaling (MDS), which *is* stable, and moreover, which typically represents more than 80% of the variation in the data. See Figure 1 for an impression and see Figure 9 (below). We have also included a stochastic version of a clustering algorithm in order to emphasize how unstable some groupings may be (Nerbonne et al. 2008).

String alignment algorithms may also be modified in many subtle ways, depending on whether one attends to base segments together with diacritics or only to base segments, whether one insists that consonants and vowels not be aligned, whether one normalizes for string length, whether diphthongs and affricates be treated as one segment or two, whether one incorporates a variable cost for substitutions depending on phonetic similarity, whether one attends to phonetic context by aligning bigrams, etc. (Heeringa et al. 2006). In this case we settled on a simple variant that is linguistically responsible, namely one in which tokenized transcriptions are used, in which consonants and vowels are always kept distinct, but in which segments are otherwise only the same or different (no variable costs), and with a normalization for word length. In keeping with our wish not to impose our view on researchers who may wish to

©Universitat de Barcelona

experiment systematically with such parameters, we allow other definitions, but not as part of the normal invocation of the analysis — modifying such parameters is a matter for experienced researchers, not beginners.

2.4 *Implementation*

This section discusses the implementation of Gabmap and may be safely skipped by readers interested only in its functionality.

Gabmap organizes users' data into PROJECTS, which each consist of exactly one map, one data set, and one measure of difference applied to the data. Separate projects must be created when users wish to work with more than one data set acquired from the same region, or with different measurements applied to the same data set. This simplifies the management of the data and also the users' views of the data. An advantage of this organization — as opposed to an organization in which the same map and data set might be associated with any number of "result" data structures, leading to a hierarchical organization — is that a number of analyses and meta-analyses (at this moment, about ten) are conducted automatically, as soon as a project is started. It is not necessary for the user to specify most options or to initialize each step in the processing chain separately. A further advantage is that the present structure allows us to add components relatively easily, a property we have already exploited.

A disadvantage of the relatively flat organization into projects is that each new sort of measurement on a given data set results in a new project. We mentioned as an advantage earlier that the user is shielded from the complexities of several analyses (and parts of analyses), but this property admittedly cuts both ways, in that it implies that the user is also unaware of many processing steps. See Figure 2 and Figure 3 for illustrations of complexity.

©Universitat de Barcelona

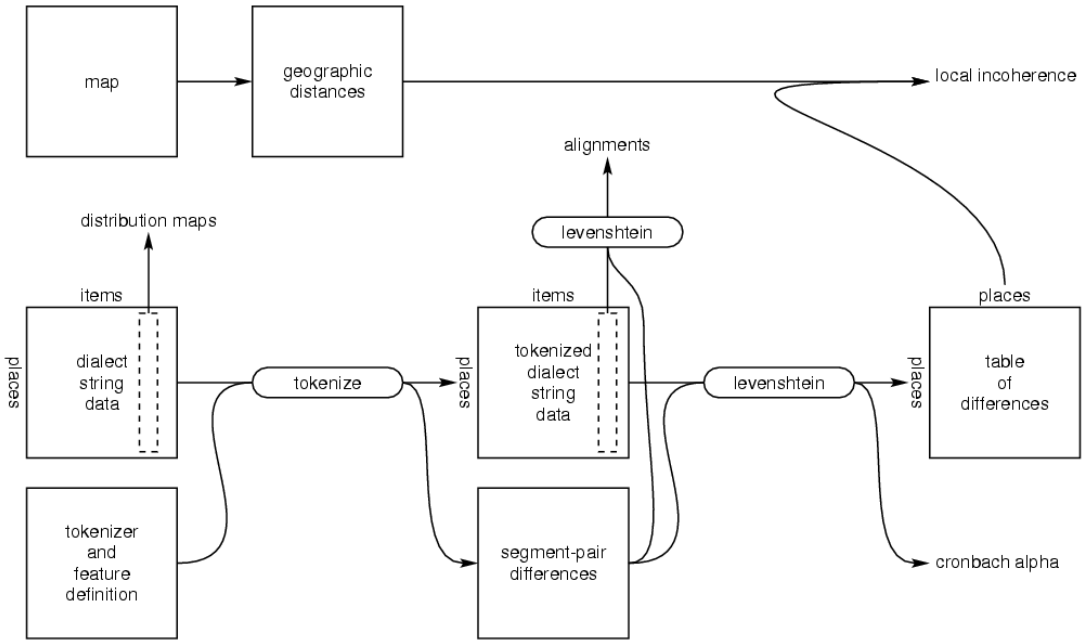

Figure 2 The flow of data at project initialization where phonetic transcriptions are compared, at which time a map and a table of "dialect string data" are input. The data is tokenized on the basis of a feature definition (which may be supplied by the user, but which is also available in a default version). Then using a table of phonetic segment differences derived from the feature definition, the alignments are made available (see "walk through" section) and word pronunciation differences are calculated. Two measures of quality are derived, Cronbach's α and "local incoherence" (Nerbonne and Kleiweg 2007). The users need to specify only a map and a data table.

Unlike many web applications, Gabmap is not built on a database, and in fact makes no use of any database whatsoever. An organization using directories and files is convenient since the programs are file based. Each user is assigned a directory with files noting login name, email address, and the like, and with subdirectories for each project. Each project directory contains a file with identifying information and general data about the project and a number of sub-directories corresponding to the results of various processing steps, e.g., sub-directories for alignments, for aggregate distances, for MDS plots, for dendrograms, and for each of a series of different sorts of maps (e.g., cluster maps, MDS projections, composite cluster maps). In addition to the results and the graphics, we store information about the processing options that led to the results.

©Universitat de Barcelona

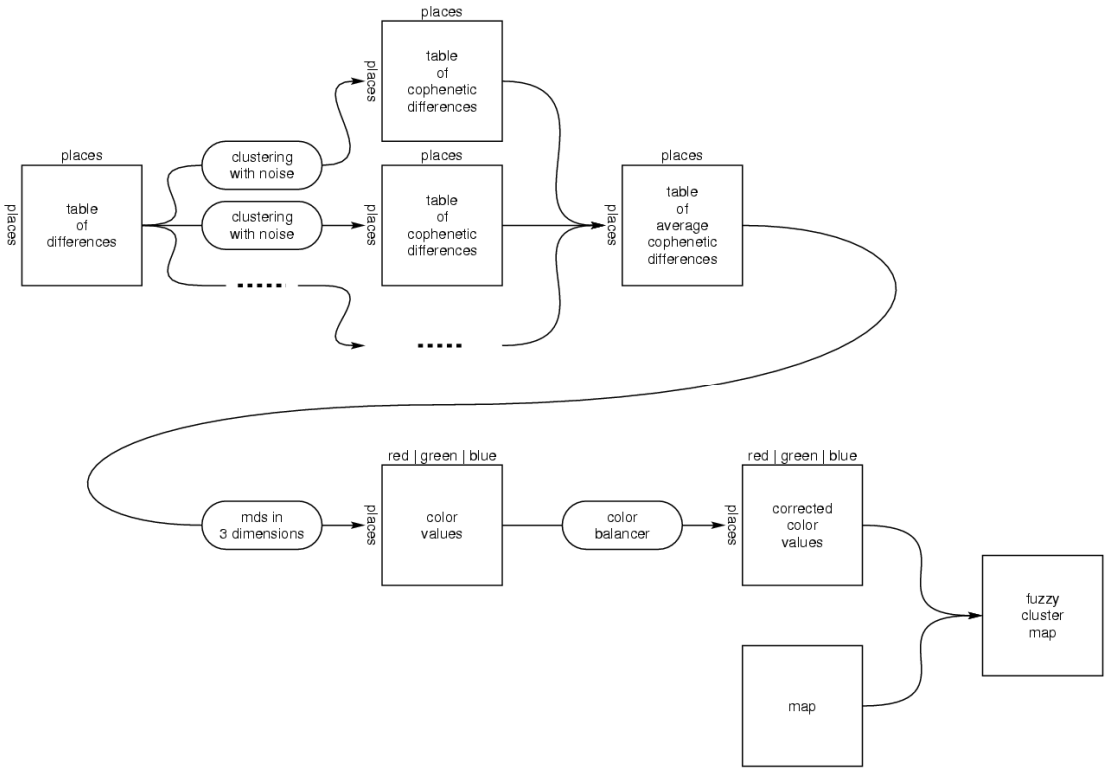

Figure 3 The flow of data in creating “fuzzy cluster maps” (see “walk through” section below for an example. This involves clustering repeatedly with random amounts of noise, while retaining for each pair of sites, their cophenetic distance (distance in the dendrogram). MDS is applied to the average distances are retained, effectively emphasizing differences more strongly, after which a map is drawn in which the first three MDS dimensions are interpreted as color intensities and projected onto the map. The user simply asks for a fuzzy cluster map.

Some of the processing steps are quite time consuming, which led us to impose a “first-in-first-out” (FIFO) discipline on user tasks. Tasks are executed serially, as they often build on one another, and because some tasks demand too much memory to be carried out comfortably in parallel. Tasks with several subtasks are not scheduled intermittently (among different users), but user feedback is provided as quickly as possible. This way users can make use of their time inspecting first (partial) results even while additional tasks are being carried out by the server. The FIFO structure can be a disadvantage when several users are working simultaneously, as it increases the waiting time of the last users in the queue considerably. In fact, since interactive requests to users result in tasks that are scheduled just as all the others, users whose jobs are processing may also experience delay. Each user must wait until all the (perhaps ten, see above) tasks of all the preceding users have all been completed. This has led us to offer tutorials to groups of fifteen to twenty participants using only smallish data sets (80

©Universitat de Barcelona

sites of 100 transcriptions each). To-date we have not experienced difficulties with this simple scheme in handling groups of this size.

Although we examined existing packages for building web applications, in particular Pylons, we were disappointed in the benefits the packages provided when compared to the additional time required to master them. Gabmap is instead implemented as a number of cgi-scripts that are invoked from within the Apache web-server. The scripts are primarily written in Python 3.1, with a brief wrapper in *sh* in order to initialize the environment. There is a special script which functions as a "dispatcher" for the different components within the application. All other interactions with the applications, e.g. the processing of specific forms, proceed via special scripts, one per interaction type. All the scripts make use of the same library of help functions in Python.

Besides Python 3.1 we made use of (i) some auxiliary scripts in Python 2.6 that cannot (now) be converted to Python 3 because they rely on libraries as yet unavailable in Python 3; (ii) external Python libraries such as pyproj (for 3.1), numpy (for 2.6 en 3.1), colormath (for 2.6); (iii) some components of the RuG/L04 software, written in Perl, and C and Flex (lexical analyser); (iv) UNIX Make and sh; (v) some components taken from the open source statistics package R; and (vi) several programs in Postscript, used not only for map-drawing, but also for calculating coordinates when users access information via "mouse-over".

2.5 *Help functions and tutorial*

A brief tutorial has been developed and presented to interested dialectologists on several occasions; and we are continually adding help functionality both in the form of immediate, brief (one-line) explanations as well as entire screens with motivation, explanation and examples. The help facilities are undergoing regular expansion.

3. Walk-through of session

To give some flavor of the web application, we present some information in the tabular or graphic form in which a user would encounter it in a Gabmap session. We

©Universitat de Barcelona

assume that the user has collected pronunciation data from Pennsylvania, USA at the 67 sites shown in the map in Figure 4 created using Gabmap's map facility. In fact we shall employ data from the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS, see Kretzschmar, 1994), and freely distributed by the US Linguistic Atlas projects.[7] We use a restricted set for simplicity.

Once the user has uploaded a data file to Gabmap, she may immediately request a list of the words whose pronunciations were elicited as well as a summary of the data noting the frequencies with which tokens such as phonetic symbols are encountered. This overview facility hones in on errors — i.e. data that could not be tokenized properly, but also on very infrequent tokens, which are also often errors in the input data. Importantly, the facility is supported by an index into the data, so that users can trace tokens back to their occurrence. It is also possible to request a map showing the frequency of a given token (or even regular expression, for the advanced), which may provide insight into its occurrence, whether this be a geographic trait, a fieldworker trait, or perhaps just a mistake. Naturally the frequency map cannot indicate definitely whether a particular transcription is an error, but users have agreed that transcriptions were in error that Gabmap highlighted because they used very infrequent phonetic symbols in ways that indicated unusual pronunciations.

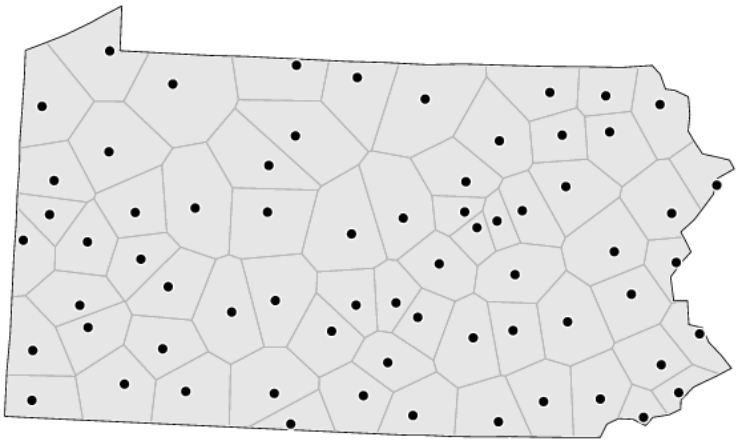

Figure 4. Map of the 67 Pennsylvania sites in LAMSAS. The map and the Voronoi tiling around the data collection sites was drawn in Gabmap.

[7] See http://us.english.uga.edu/

©Universitat de Barcelona

The initial result of comparison is a site × site distance table, in which half the values simply repeat (due to the symmetry of the distance measures). But this still leads to 2,211 cells with distance values even in our artificially small set – surely too many for direct inspection. For this reason Gabmap supports the further analysis with appropriate visualizations and statistical analyses. Beam maps (Goebl's *Strahlenkarten*) provide an excellent aggregate view of the data. In principle a line is drawn between each two sites where the darker the line, the more linguistically similar the sites. Particularly coherent areas are normally immediately visible as dark collections, and boundaries appear as lighter-colored swaths. See Figure 5. Network maps connect only adjacent sites, again coloring more darkly in case the sites are linguistically similar. They offer a less complete, but also a clearer illustration of the linguistic differences measured.

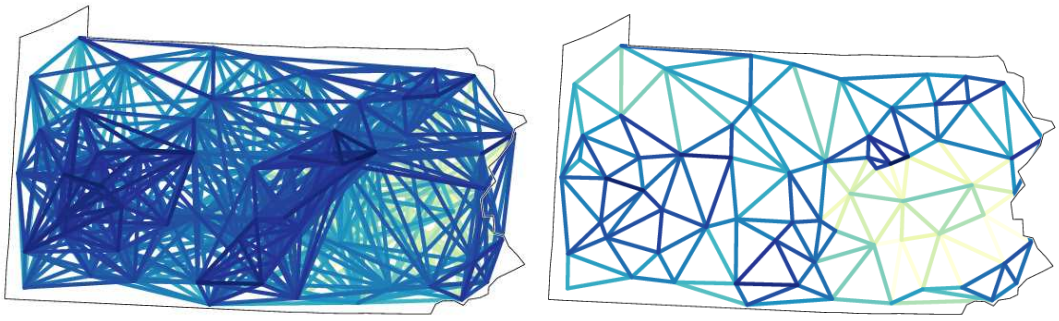

Figure 5. A beam map (left) and a network map (right) displaying the pronunciation differences measured in a fairly straightforward way.

Naturally researchers will wish to check data measurements in various ways, e.g. by comparing the calculation of differences. Gabmap supports such wishes, e.g. by displaying the alignments used when a user requests this (by specifying a word whose alignments the user wishes to see). Note that even in the small data set we are examining here, there are 2,211 pairs of sites with 150 pronunciation comparisons per site — over 300,000 in total. In principle, a research may examine any of these. See Figure 6 for an example.

©Universitat de Barcelona

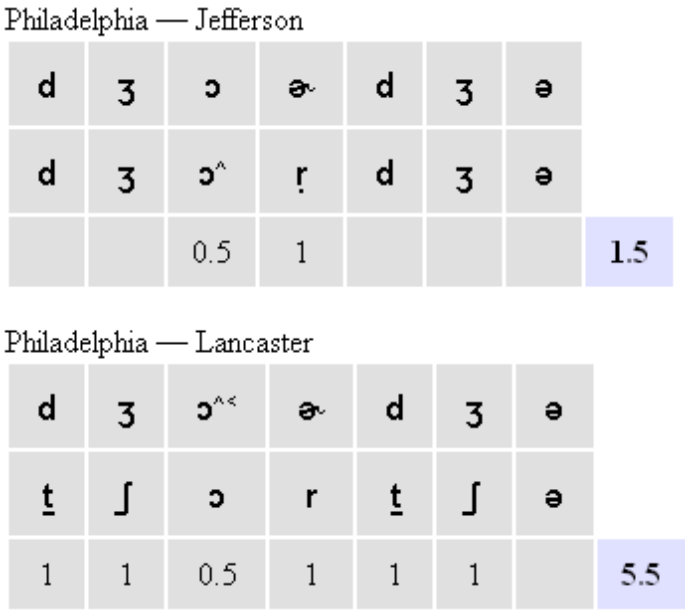

Figure 6. An example alignment and pronunciation distance calculation from Gabmap, showing pronunciations of the state name 'Georgia' at three sites in Pennsylvania. Note that Gabmap is not restricted to simple Unicode or X-SAMPA encodings of IPA transcriptions.

Of course a central topic in traditional dialectology (of the "German" school, see Kretzschmar, 2006) has been the determination of dialect boundaries, and Gabmap supports this using clustering of four sorts. The determination of boundaries using clustering is an alternative to the isogloss techniques noted in the discussion above (on Chambers and Trudgill's remarks) in that the boundaries determined need not correspond to any single isogloss, but only to the aggregate difference. Naturally, researchers are interested not only in the dendrogram tracing the history of the clustering procedure, but also in its projection to the map. Figure 7 shows both, where the colors in the map and dendrogram are linked.

©Universitat de Barcelona

Figure 7. A dendrogram resulting from clustering the aggregate distance table as well as its projection to the map of Pennsylvania. Note the colors are linked for easy reference. Those interested in American dialectology may recognize Kurath and McDavid's famous "Route 40" boundary stretching east to west across the northern part of the state (Kurath 1949; Kurath & McDavid 1961).

While the determination of dialect areas is important in comparing quantitative work to traditional dialectometry, contemporary techniques rely on clustering, which is less than 100% reliable. It is essential therefore to compare clustering results such as the one in Figure 7 to more reliable statistic analyses, such as multi-dimensional scaling (MDS) (Embleton, 1993). Gabmap has introduced a special "cluster validation" module for this purpose. MDS is applied to the aggregate distance table with the result that each site is assigned coordinates in the plane in the manner suggested by Figure 8.

©Universitat de Barcelona

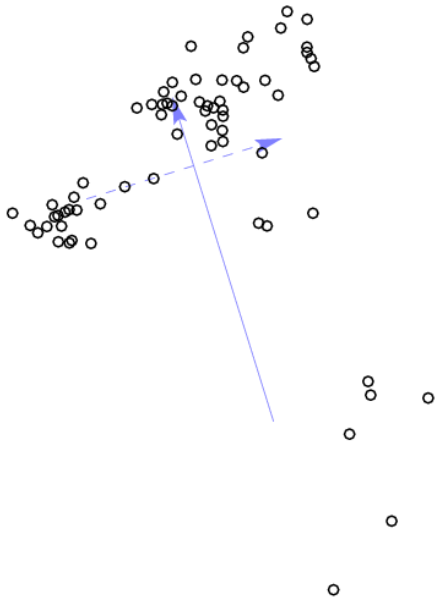

Figure 8. A two-dimensional MDS analysis of the aggregate pronunciation distances in Pennsylvania. The distances implicit in this scatter plot (as measured by a ruler) correlate very highly with the distances in the original aggregate linguistic distance table (r=0.94).

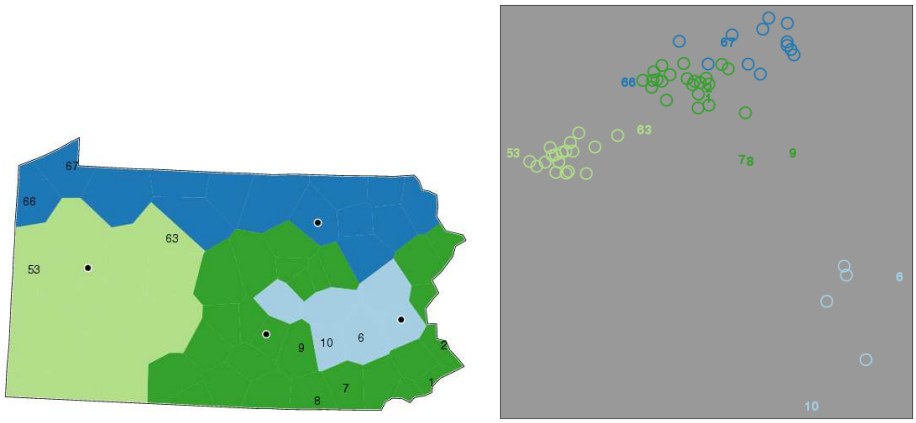

Figure 9. A novel facility in Gabmap allows researchers to compare the cartographic projection of clusterings (left) with MDS results (right). The MDS plots shows us at glance that the light blue area (around Franklin county) is quite distinct, but also quite diverse, that the southwestern area is nicely distinct, but that the dark green and dark blue areas appear not to be discriminated well, and deserve closer attention. The validation facility allows the user to examine just the two areas in question, a more sensitive view. The numbers on the map (left) re-appear in the MDS plot (right) to enable the researcher to identify sites that do not fit nicely in the clusters.

There are also attempts to remedy the inherent instability in clustering by adding stochastic elements to the process, notably the bootstrap (Felsenstein, 2004: Chap.20), which involves repeatedly re-sampling the set of words repeated to obtain a new

©Universitat de Barcelona

sample, and in particular allowing the same word to appear more than once in the new re-sample. The results of the repeated clustering tend then to be more stable. Due to the computational cost of the bootstrap, we use an alternative procedure which has been shown to correlate nearly perfectly (Nerbonne et al. 2008) in which small amounts of random noise are added to different clustering analyses of the same distance table. The result is a PROBABILISTIC DENDROGRAM in which the groups are assigned a confidence level, namely the number of times the group emerged in the "noisy" repetitions. Figure 10 shows an example together with its projection to the geographic map.

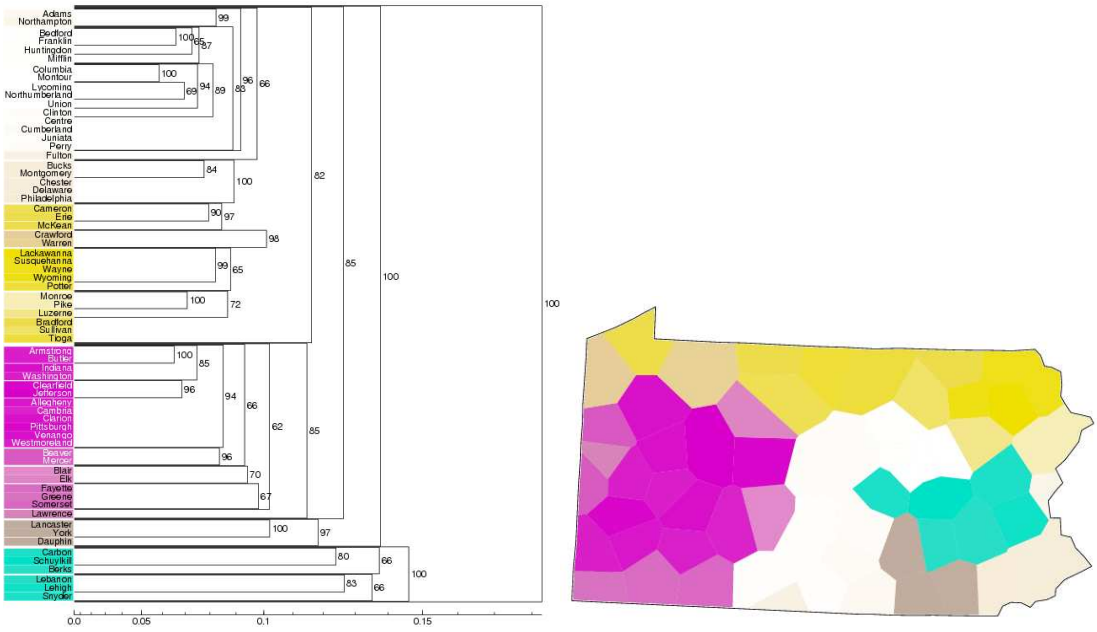

Figure 10. Noisy clustering correlates highly with clustering using the bootstrap (Felsenstein 2004: Chap. 20) and assigns a confidence to each group (the small numbers to the right of the brackets indicate the percentage of "noisy" clusterings in which the bracketed group was found. The results (after applying MDS to the branch lengths of the dendrogram on the left) may also be projected to a map (right).

It is clear that linguists are more interested in the details of dialect distributions than in the aggregate. From the point of view of linguistic theory, one is interested not in the fact that there is a fairly coherent dialect area in southeastern Pennsylvania (around Schuykill, Berks and Lehigh), but what linguistic features are responsible for the differences. Gabmap therefore supports users' search for such features. We emphasize that the researcher is free to examine any number of distributions — both distributions of individual pronunciations of words, but also distributions of more abstract patterns which might be specified by regular expressions. Figure 11 provides an

©Universitat de Barcelona

example of one such search. Examining the pronunciations of the word ‘Georgia’ in Pennsylvania, the researcher has asked to see the geographical distribution of the words in which the initial voiced affricate [dʒ] is pronounced voicelessly, as [tʃ]. Although this example involved an area that was identified via clustering, researchers are free to examine the geographic distribution of any feature.

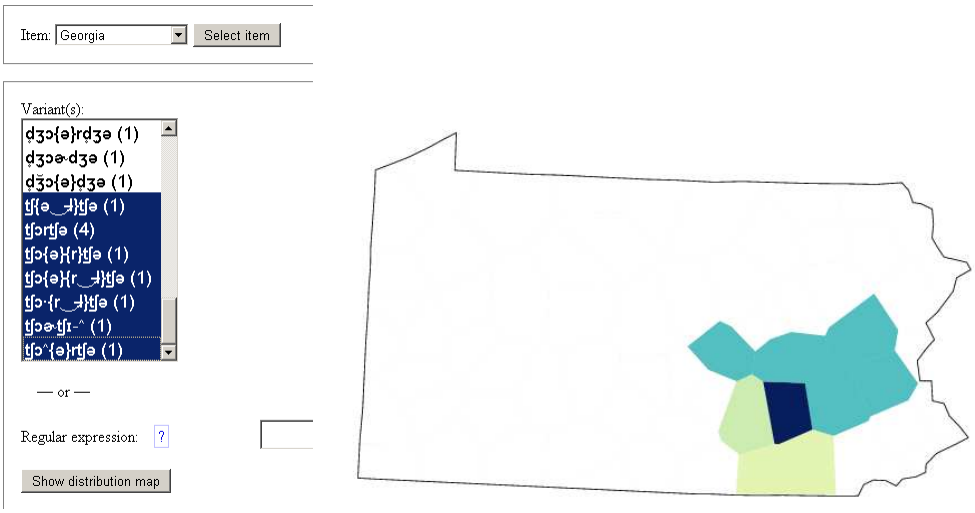

Figure 11. A distribution map for the pronunciation of ‘Georgia’ using an initial voiceless affricate [tʃ]. In fact, researchers are free to examine the distribution of any set of pronunciations (of a single item) they like, and even to specify a more abstract pattern via a regular expression. This facility does not depend on first obtaining the results of aggregate analysis and is therefore of broader interest to dialectologists.

Gabmap supports linguistically oriented research in other ways as well. Given the results of an aggregate analysis, it is also interesting to examine pronunciation variants to attempt to determine the identifying features. Wieling & Nerbonne (to appear) suggest two quantitative measures of the degree to which a feature identifies a dialect area, its REPRESENTATIVENESS and its DISTINCTIVENESS. The degree to which a feature is representative of a dialect area is the fraction of sites in the area at which it may be found (beyond a threshold level). And a feature in a dialect area is distinctive with respect to the larger language area to the degree that it occurs exclusively in that area. See Wieling and Nerbonne (to appear) for the formulas. Figure 12 shows the information provided to the researcher about the pronunciation of ‘miles’ ([maɪls]) in the southeastern Pennsylvania area where German was often spoken until the early twentieth century. Note that the plural is pronounced [s] and not [z] as in standard American pronunciation, and also that the [l] is not velarized, as it often is elsewhere.

©Universitat de Barcelona

By employing measures of the identifying quality of features we try to lead the researcher to insightful candidates.

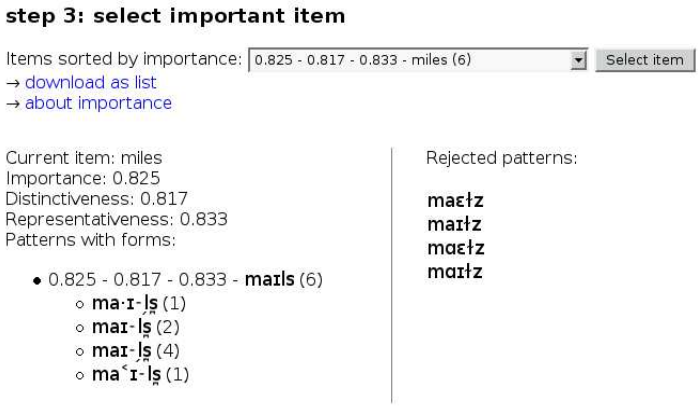

Figure 12. Gabmap includes a facility to aid the linguistically oriented researcher in searching for features that might identify a candidate dialect area.

Finally, Gabmap also provides some aggregate statistics and graphs showing the distribution of linguistic variation. Figure 13 shows a graph plotting aggregate linguistic distance as a function of distance. A local regression line is drawn as well as a logarithmic one. The local line is drawn to give a sense of the degree to which the logarithmic line represents the data well. Nerbonne (2010) discusses this sort of distribution in more detail.

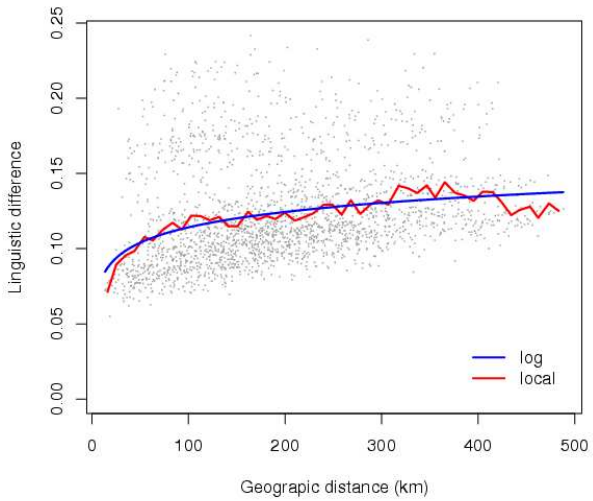

Figure 13. Gabmap provides facilities for examining the distribution of aggregate variation as a function of geography. The local regression line is also drawn is suggest how well the logarithmic line fits the data.

©Universitat de Barcelona

4. Reactions from users

Gabmap has been presented at a number of workshops and invited tutorial sessions. The feedback has been quite positive. Users familiar with the older RuG/L04 software have been very happy to hear that there is a more user-friendly graphical user interface available. Researchers from other fields within dialectology (other than dialectometry) have been especially enthusiastic to hear about the tool for creating distribution maps available in Gabmap, since a free and easy-to-use tool for distribution maps has not previously been available. Of the dialectometric analyses offered in Gabmap, cluster analysis seems to be the most appealing one to users. It has been important to emphasize to researchers who are new in the field that cluster analysis is not a particularly stable method, and that the methods for cluster validation offered in Gabmap should be applied. Some users have made suggestions for further development of Gabmap. A wish from several research groups has been to make it possible to define potential dialect areas manually; at the moment only one geographic coordinate can be supplied for each data collection point, and the area surrounding each data point on the map is computed automatically using Voronoi tiling. A further wish has been to implement more options for distribution maps.

5. Discussion. Future Ideas

We believe that Gabmap has the potential to lower the technical threshold to dialectometry enough to stimulate exploration and criticism. As proponents of dialectometry we are most interested in the former, i.e. stimulating the broader use of quantitative techniques in dialectology, but we would also welcome the latter, i.e. a better informed criticism of dialectometry. This is also normally a productive scientific path.

We see many further opportunities for Gabmap and related services. First, there are points in the present Gabmap we would prefer to see changed. The most significant of these is the absence of output in the form of geo-referenced maps. The maps produced by Gabmap are attractive and insightful, but we should like to superimpose them on other maps easily in order to compare distributions and boundaries of dialect

©Universitat de Barcelona

phenomena with those of other phenomena such as trade, communication and popular culture. A second, less significant shortcoming we are even now working to eliminate is the sparseness of the documentation. A third problem is the organization of work into projects (described) above. While we do not have a plan for improving this, we concede that users find it counterintuitive that they cannot try different measuring techniques within a single project.

There are also opportunities for further development. Probably the most important of these would involve making it easier for others to contribute modules, i.e. adopting an open-source development mode. Once it becomes easier for others to contribute, then the scientific imagination is the limiting factor. Further suggestions for improvement have been contributed by users (see Section 4 above).

Acknowledgments

We thank the CLARIN-NL program (see footnote 1 above) and also Jan Pieter Kunst, who piloted Gabmap onto a CLARIN-compliant server at the Meertens institute. Wilbert Heeringa (Meertens Institute, Amsterdam) used Gabmap in a course on dialectology he taught at the University of Amsterdam in Sept. 2010; he and his students provided many specific comments and suggestions. Martijn Wieling (Groningen) assisted both Simonetta Montemagni (Institute for Computational Linguistics, *Consiglio Nazionale delle Ricerche,* Pisa) and Lucija Simicic (Institute for Research in Anthropology, Zagreb), who conducted important first studies using Gabmap and were generous in providing feedback. Sebastian Kürschner (Erlangen) invited Therese Leinonen to conduct a tutorial on Gabmap at the *Tagung des Forums Sprachvariation* of the *Internationale Gesellschaft für Dialektologie des Deutschen e.V.* (IGDD) (see http://www.igdd.gwdg.de/) in Erlangen on Oct. 15, 2010. An anonymous reviewer was helpful in criticizing an early draft.

©Universitat de Barcelona

References

AURREKOETXEA, Gotzon & Jose Luis ORMAETXEA (eds.) (2010) *Tools for Linguistic Variation*. Supplements of the *Anuario de Filología Vasca* "Julio Urquijo", LIII, Bilbao: University of the Basque Country.

BLOOMFIELD, Leonard (1927) Language, New York: Holt, Rhinehart and Winston.

CHAMBERS, J.K. & Peter TRUDGILL (1998) *Dialectology* (2nd ed.), Cambridge: Cambridge University Press.

EMBLETON, Sheila (1993) "Multidimensional Scaling as a Dialectometrical Technique: Outline of a Research Project", in Reinhard KÖHLER & Burghard RIEGER (eds.), *Contributions to Quantitative Linguistics*, Dordrecht: Kluwer. 267-276.

FELSENSTEIN, Joseph (2004) Inferring Phylogenies, Sinauer, Sunderland, MA.

GOEBL, Hans (2006) "Recent Advances in Salzburg Dialectometry", *Literary and Linguistic Computing* 21(4):411-436, Spec. Iss., *Progress in Dialectometry: Toward Explanation* ed. by John NERBONNE & William KRETZSCHMAR, Jr.

GOEBL, Hans (2010) "Dialectometry: Theoretical pre-requisites, practical problems and concrete applications (mainly with examples from the Atlas Linguistique de la France 1902-1910)", *Dialectologia*, Special issue, I, Geolinguistics around the World, 63-77. <http://www.publicacions.ub.edu/revistes/dialectologiaSP2010/>

HAIMERL, Edgar (1998) "A Database Application for the Generation of Phonetic Atlas Maps", in John NERBONNE (ed.), *Linguistic Databases*, Stanford: CSLI Press, 103-116.

HAIMERL, Edgar (2006) "Database Design and Technical Solutions for the Management, Calculation and Visualization of Dialect Mass Data", *Literary and Linguistic Computing* 21(4): 436-444, Spec. Iss., *Progress in Dialectometry: Toward Explanation* ed. by John NERBONNE & William KRETZSCHMAR, Jr.

HEERINGA, Wilbert, Peter KLEIWEG, Charlotte GOOSKENS & John NERBONNE (2006) "Evaluation of String Distance Algorithms for Dialectology", in: John NERBONNE & Erhard HINRICHS (eds.) *Linguistic Distances*, Workshop at the joint ACL-COLING, Sydney. 51-62.

KLEINBERG, Jon (2003) "An Impossibility Theorem for Clustering" In: Suzanna BECKER, Sebastian THRUN & Klaus OBERMEIER (eds.) *Advances in Neural Information Processing Systems* 15, Cambridge: MIT Press. Available online at http://books.nips.cc/papers/files/nips15/LT17.pdf.

KLOEKE, G.G. (1927) De hollandsche Expansie, The Hague: Martinus Nijhoff.

KRETZSCHMAR, William, Jr. (ed.) (1994) *The Handbook of the Linguistic Atlas of the Middle and South Atlantic States*, Chicago: The University of Chicago Press.

©Universitat de Barcelona

KRETZSCHMAR, William, Jr. (2006) "Art and Science in Computational Dialectometry", *Literary and Linguistic Computing*, 21(4), 499-410. Spec. Iss., *Progress in Dialectometry: Toward Explanation* ed. by John NERBONNE & William KRETZSCHMAR, Jr.

KURATH, Hans (1949) *A Word Geography of the Eastern United States*, Ann Arbor: University of Michigan Press.

KURATH, Hans & Raven MCDAVID (1961) *The Pronunciation of English in the Atlantic States*: *Based on the Collections of the Linguistic Atlas of the Eastern United States*, Ann Arbor: University of Michigan Press.

NERBONNE, John (2009) "Data-Driven Dialectology", *Language and Linguistics Compass*, 3(1), 175-198.

NERBONNE, John (2010) "Measuring the Diffusion of Linguistic Change", *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365:3821-3828. DOI: 10.1098/rstb.2010.0048.

NERBONNE, John & Wilbert HEERINGA (2010) "Measuring Dialect Differences", in Jürgen Erich SCHMIDT & Peter AUER (eds.), *Language and Space: Theories and Methods*, Berlin: Mouton De Gruyter, 550-567.

NERBONNE, John, Peter KLEIWEG (2007) "Toward a Dialectological Yardstick", *Journal of Quantitative Linguistics*, 14(2), 148-167.

NERBONNE, John, Peter KLEIWEG, Wilbert HEERINGA & Franz MANNI (2008) "Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering", in Christine PREISACH, Lars SCHMIDT-THIEME, Hans BURKHARDT & Reinhold DECKER (eds.), *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society*, Berlin: Springer, 647-654.

NERBONNE, John, Jelena PROKIĆ, Martijn WIELING, & Charlotte GOOSKENS (2010) "Some Further Dialectometrical Steps", in AURREKOETXEA, Gotzon & Jose Luis ORMAETXEA (eds.), *Tools for Linguistic Variation*. Supplements of the *Anuario de Filología Vasca* "Julio Urquijo", LIII, Bilbao: University of the Basque Country, 41-56.

PROKIĆ Jelena & John NERBONNE (2008) "Recognizing Groups among Dialects", *International Journal of Humanities and Arts Computing*, 2(1-2), 153-172. Special issue on *Computing and Language Variation*, ed. by John NERBONNE, Charlotte GOOSKENS, Sebastian KÜRSCHNER & Renée VAN BEZOOIJEN. DOI: 10.13366/E1753854809000366.

WIELING, Martijn & John NERBONNE (to appear) "Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features", *Computer Speech and Language*. DOI: 10.1016/j.csl.2010.05.004.

Available online (http://www.sciencedirect.com/) since May 21, 2010.

©Universitat de Barcelona