Dialectologia 17 *(2016), 1-22.*
ISSN: 2013-2247

Received 15 March 2015.

Accepted 21 July 2015.

DIATECH: TOOL FOR MAKING DIALECTOMETRY EASIER

Gotzon AURREKOETXEA, Gotzon SANTANDER, Iker USOBIAGA & AITOR IGLESIAS

University of the Basque Contry[1]*

gotzonaurre@gmail.com/sunnyworld84@hotmail.com/ikeruso@gmail.com/

txadonak@gmail.com

Abstract

Diatech[1] is a web application to analyze linguistic differences in a quantitative exploration in a friendly way, allowing users even if they have not computational expertise. The application conceived as a tool which provides all tools that a dialectologist needs when his objective is to draw conceptual maps (the occurrence of individual features), to delimitate the linguistic distance between dialects or to draw boundaries in dialect areas. *Diatech* creates different types of maps (isoglotic, beam, similarity maps, etc.), cluster and MDS analyses, it checks out centroid localities and analyzes the linguistic features that provoke the main variation. It uses different linguistic measures as RIV (Relative Index Value) or Levenshtein algorithm, taking into account different types of answers (such as orthographic, phonetic or lemmas). All maps and illustrations can be downloaded in different image format (RGB o CMYK) and sizes. Since the application was launched, the responsible team has continued improving it, making it easier to use and more powerful in the statistic techniques.

Keywords

linguistic variation, dialectometry, computer program, automated maps

* UPV/EHU, Irakaslegoaren Unibertsitate Eskola, Leioa, Bizkaia.

[1] Diatech is accessible at http://eudia.ehu.eus/diatech.

©Universitat de Barcelona

DIATECH: UNA HERRAMIENTA PARA UNA DIALECTOMETRÍA MAS ASEQUIBLE

Resumen

Diatech es una aplicación web para el análisis de las diferencias lingüísticas desde el punto de vista cuantitativo en un entorno amigable, para uso de lingüistas no habituados a usar programas informáticos sofisticados. La aplicación ha sido concebida como una herramienta que proporciona todos los útiles necesarios para el dialectólogo que pretende crear mapas conceptuales (las ocurrencias de características individuales), delimitar distancias lingüísticas entre dialectos o determinar las fronteras de áreas dialectales. *Diatech* puede crear diferentes tipos de mapas (mapas isoglóticos, mapas de rayos, mapas de similitud, etc.), análisis de clúster y de escalamiento multidimensional; puede también determinar las localidades centroides de un área dialectal o detectar las características lingüísticas que provocan la mayor variación. La herramienta usa diferentes medidas lingüísticas, como RIV (Relative Index Value) o la distancia de Levenshtein, teniendo en cuenta diversos tipos de respuestas (ortográficas, fonéticas o lemas). Tanto los mapas como las diversas ilustraciones pueden ser descargados en diferentes formatos (RGB o CMYK) y tamaños. El equipo continúa mejorando e implementando nuevos recursos estadísticos con el objetivo de hacer la herramienta más accesible a los investigadores.

Palabras clave

variación lingüística, dialectometría, programa informático, cartografía automatizada

1. Introduction and background

1.1 Background

The study of dialect boundaries has been carried out since the beginning of dialectology. Fortunately, the discussion about the existence of the boundaries it is over and nowadays it is accepted by all dialectologists. The study about these boundaries has been carried out by the isogloss methods in the traditional dialectology. Nevertheless, this method has been criticized (Kessler 1995, Goossens 1977, Inoue 1996, among others), while other methods have been applied in the last decades.

By quantifying linguistic data and using statistical procedures, modern dialectology has been able to measure linguistic distances between different localities and classify dialects in a more accurate and scientific way. Besides its suitability to measure

©Universitat de Barcelona

differences between dialects, dialectometry shows the analysis of the dialects in a much more attractive way than it use to do in the past.

Jean Séguy (1973) was the creator of the discipline, although he had to deal with quantification of the data in a manual way, by making by hand all quantification task. He was able to measure linguistic distances between close and far localities in a bidirectional way, taking two localities each time and finally drawing the Gascony's features map. In contemporary works such as H. Goebl (1976) and H. Guiter (1973), they started working in a similar way. But it was Goebl who really developed and achieved an enormous progress. And he was the one who conceived the first computing package for quantitative dialectology. This package, named VDM and created by Haimerl (1999), introduced all techniques that Goebl had been developing in the last 30 years. It uses different linguistic distances (RIV and WIV), wide variety of synoptic maps (similarity, transitional and conservative areas, etc.), beam and isoglotic maps, cluster and correlation analysis (Goebl 2010). This desktop program gathers up the tools that a dialectologist needs from the geolinguistic point of view, facilitating users to draw individual features maps and synthetic maps, using quantitative analyses for achieving the mentioned purpose.

The use of computerized statistic packages compels users to learn not only Linguistics but also Statistics, because dialectometry is a multidisciplinary task, in which dialectal data must been analyzed in a quantitative way too.

The second package in dialectometry was developed by Peter Kleiweg's L04 package (www.let.rug.nl/kleiweg/L04). This package was finally developed in the Gabmap online version (Nerbonne *et al.* 2011).

1.2 Motivation

Although we knew the computerized programs for developing dialectometry, we began with the task of creating a new one, because we had two concerns in mind, that they have not been accurately solved yet by those sorts of software: on the one hand, the problem of "multiple responses" and, on the other one, the transportability of the statistical outcomes (Aurrekoetxea *et al.* 2013).

©Universitat de Barcelona

The "problem" of the multiple responses (henceforth MR) is relatively new in dialectology. The traditional dialect atlases do not gather more than one response for each question from one locality (mainly they gather the "best" response, when more than one is collected). Nevertheless, every linguistic knows that language is constantly changing; whether one of the main features of the language is to be changing continuously, in the real situation of any of them, will be cases that show this variation.

Whether the traditional dialect atlases do not gather this intralocality variation is a cause of the used methodology. This subject is changing in modern atlases, in which linguistics use different methodologies, taking into account the need of recording this local variation. This is the case, for example, in the Linguistic Atlas of Basque (henceforth EHHA) (Euskaltzaindia 2010-2016). The treatment of these MRs in the quantitative dialectology has been a constant concern (Aurrekoetxea 2002). The VDM package did not consider this kind of data, and we disagreed with the solution given by Gabmap (Aurrekoetxea *et al.* 2013: 26-27).

The second motivation, the transportability of the statistical outcomes, is more complex and requires improves from theoretical and statistical points of view. In science it is crucial the transportability of the outcomes from one team to another; it is essential for the advance of the science. However, it is not the case of dialectology. Dialectologists have epistemological weakness when some basic concepts have to be defined. Even if there is an amount of definitions about 'dialect', each dialectologist uses it in different ways. It is a clear consequence of the lack of accuracy in the definition of this term. The fact that there are many definitions is an evidence of the lack of accuracy. Now, having improved in the measurement of the linguistic distance or similarities and being accustomed to statistical techniques which allow us to measure these distances in a much more accurate way than with the tools of the traditional dialectology, dialectologists must improve in the data quality and quantity that has to be used.

Taking into account these reasons, we have put our efforts into creating a new package for doing dialectology. Concretely, a software that will continue improving the performance. This version, firstly, provides tools for uploading data to the internet and for drawing the corresponding conceptual maps; secondly, it offers the possibility of managing databases which are hosted into it and of carrying out many dialectometrical

©Universitat de Barcelona

analyses (similarity maps, transition zones, detection of conservative zones, analysing the correlation between two fields, beam and honeycomb maps, — deterministic and probabilistic-cluster analysis —, MDS, centroid localities and the analysis of the linguistic features that provoke the main variation), using string (Levenshtein) and nominal or categorical (RIV (Relative Index Value-RIV) and WIV (Weighted Index Value-WIV)) linguistic distances following different DM schools and taking into account different types of answers (such as orthographic, phonetic or lemmas — Goebl's *taxat*).

2. The main features of *Diatech*

Diatech is based on the main dialectometrical methods developed until nowadays and includes the most useful tools that are described in the scientific literature. Apart from the tools to manage the project (importing and exporting of the data, management of the database (questions, answers, locations and informants) and project, the tool provides a wide range of technical procedures for doing DM, allowing users to carry it out in different steps: selecting the type of answer, linguistic field, linguistic distance, algorithms and analysis type.

2.1 Uploading data to Diatech

The first step is one of the most difficult ones to be overcome by users who are not skilful in the use of computers. In fact, it is very easy, but linguists must know that it is enough to insert one extra comma by mistake in the database or to leave one out and we will not be able to upload the data. Taking into account these difficulties, the team of the tool has decided to help the user by including only one database structure. This database must have three kinds of data: locations, questions and answers.

©Universitat de Barcelona

	location1	location2	location3	...
question1	answer1-locat1	answer1-locat2	answer1-locat3	...
question2	answer2-locat1	answer2-locat2	answer3-locat3	...
question3	answer3-local1	answer3-locat2	answer3-locat3	...
...	...	...	...	...

Figure 1. Structure of the database

The structure of the single file contains these features: the first line should include the locations. In the following lines, the first column should have the translated questions with their corresponding language and the other columns should have the answers as it is shown in Figure 1. Whether there is more than one answer in a locality, the answers will be separated by commas.

In some geolinguistic projects localities can have more than one informant; whether users would like to analyze and compare the data of different informants, they have to create as many databases as informants and upload them separately, but it is different if informants have not sociolinguistic motivations (age, social class, etc.) and has not sense to distinguish the responses of them, answers of different informants must be stored in the same database. That is to say, if the user would need to analyze separately the data of two generations, or data of urban and not urban informants, he must create two databases and upload them to *Diatech* in two separate projects.

The requirements for the data structure are the following ones: it must be coded as UTF8 to avoid problems with special characters; it must be stored as .csv format; and finally, they have to be compressed all together in one ZIP file.

2.2 Creating the map

Once the data has been uploaded, we must create the map, including in it all the localities that we have in the database and drawing the boundary. In order to do this, the *Diatech* tool provides a Google map in which the tool has located the localities. Nevertheless, the name of one locality may sometimes be situated in different parts of the world; in these cases, the user has to deal with it, managing and putting it in the

©Universitat de Barcelona

correct position on the map. Once the marking of the boundary is completed, the machine will automatically create the Voronoi Map (Thiessen polygons) including all the localities of the project. Once this is done, we have completed the storage of our data in *Diatech.*

2.3 Managing the project and data

When the data is uploaded, users can manage not only the data, but also all the features of the project. We have different options for using the tool: to manage the project (invite users to the project, export the database, etc.), to manage the database (changing, correcting, adding or deleting locations, questions and answers), to search for answers (by questions, location, etc.), or to display data in maps. Here, we will consider only three aspects: questions, the linguistic field and the type of answer.

2.3.1 Questions

If we chose "question" we will have the list of all the questions of the database, in the same order as in our previous database system, and beginning with the language in which the questions are written.

2.3.2 Linguistic domain or linguistic field

As far as the "linguistic domain" is concerned, the user has different options: if he has data with more than one linguistic domain or field, he can group them into different projects, such as phonology, noun or verb morphology, syntax, prosody and lexicon; this means that each one can be analysed separately, in different projects, to study each field and finally all the projects can be analyzed together, to measure the total linguistic distance.

©Universitat de Barcelona

2.3.3 Type of answer

Concerning the "type of answer", the *Diatech* tool allows people to use different transcription systems of data: phonetic transcription, orthographic transcription or lemma. If the researcher would like to measure linguistic differentiation with the best possible accuracy he must use the phonetic level of transcription system. But he has the option to use less accurate transcription by using orthographic transcription. Finally, the researcher has also the option to use *lemmas*, to take only the essential features, avoiding accidental differences that could be embedded into the data; for example, lexical data can have features that are not strictly belonging to the lexical field, such as phonetic or pronunciation features, and so on. Each type of answer must be stored and uploaded to *Diatech* in a different project.

2.4 Multiple responses (MR)

Diatech gives a suitable solution to the MR problem in dialectology in order to measure multiple responses, similarity coefficients are calculated, particularly the Dice coefficient. Taken two linguistical identities a and b where the respective set of responses have been A and B the similarity index between the two of them would be defined has:

$$\frac{2|A \cap B|}{|A|+|B|}$$

where $|A|$ is the amount of responses gathered in a, $|B|$ the amount of answers in b and $|A \cap B|$ the amount of answers in common in both locations.

If a dissimilarity distribution is requested the complementary formula applies:

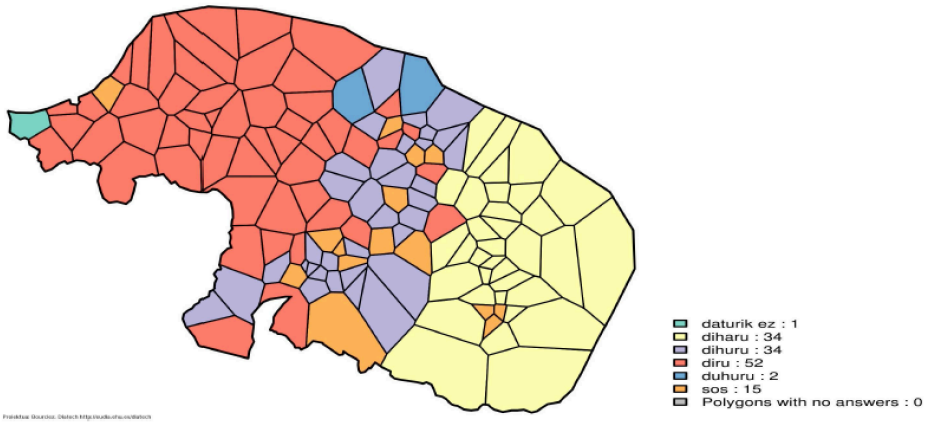
$$1-\frac{2|A \cap B|}{|A|+|B|}$$

©Universitat de Barcelona

Dialectologia 17 *(2016), 1-22.*
ISSN: 2013-2247

2.5 Conceptual maps

Once the database is uploaded and the map is created, the user can draw conceptual maps, as if it was a linguistic atlas; it can be made as many maps as questions have been made. All maps provided in *Diatech* are drawn by using Voronoi polygonation. The conceptual maps, drawn in colours, will be furnished with legend, in which the response will be provided with a determined colour. In Map 1 it is shown the conceptual map referring to 'money' with Bourciez corpus data (Aurrekoetxea, Videgain & Iglesias 2007) (see Map 1). You can see that there are five responses *(diharu, dihuru, duhuru, diru* and *sos*), each one with its colour. The map shows the distribution of each word.



Map 1. 'Money' conceptual map (Bourciez corpus)

3. Doing Dialectometry (DM)

Three kinds of DM can be made by using *Diatech:* unidimensional DM, multivariate aggregate DM and correlative DM. To do all these quantitative analyses, users have to select some basic features as the linguistic distance and the tool provides different algorithms.

©Universitat de Barcelona

3.1 Basic features

Apart from the data type already seen, the tool provides a wide range of types of analysis, linguistic distances and algorithms. *Diatech* supplies the most frequently used similarity measures or linguistic distances in DM; that is nominal or categorical and it is a string measurement. When the analysis is made with lemmas we propose to use categorial measures; among these measurements *Diatech* uses RIV 'Relative identity Value' and WIV 'Weighted Identity Value' (Goebl 1992, 2007, 2010), used by *VDM.* But when the quantitative analysis is made with phonetic or orthographic answers we propose to employ 'Levenshtein distance' (see Nerbonne & Heeringa 2001; Heeringa, Nerbonne & Spruit 2007: 5; Heeringa 2004, for more details), used by *Gabmap*. The researcher has the option to choose one of these, according to the aim of his research. There are more distances: Euclidean, Manhattan distance (Prokic 2009), etc., but there are not widely used in DM and that is why *Diatech* does not provide them.

For the visualisation of the results, *Diatech* gives users the option to use three algorithms: *Med*, *MinMwMax* and *MedMw* (for more information about these algorithms see Goebl 2013).

3.2 Unidimensional DM

There are three main ways to analyse data by starting with one locality and comparing data from two localities each time: synoptic maps, honeycomb maps and beam maps.

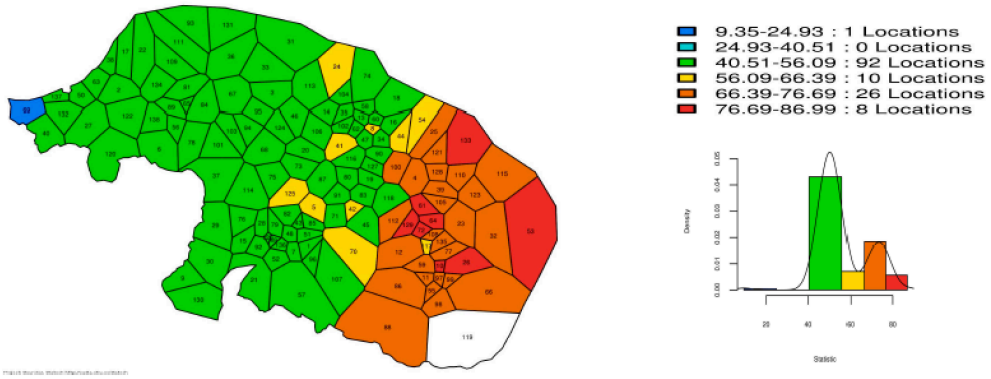
3.2.1 Synoptic maps

The synoptic map in DM is a synthetic map in which the researcher shows the linguistic similarities / dissimilarities among the localities. There are many kinds of synoptic maps: maps of distribution of similarity, maps to show the transitional zones, maps to show conservative areas, maps to show the linguistic center of the dialectal area, etc.

©Universitat de Barcelona

3.2.1.1 Distribution of similarity

Departing always from one locality, the distribution of similarity shows the linguistic difference of each locality respect to the other. The distribution of similarity can change depending on the type of data, the algorithm and the linguistic measure that has been used. Check, for example, the distribution of similarity of Santa Garazi (Map 2), using the *MinMwMax* algorithm.



Map 2. Distribution of similarity from Santa Garazi

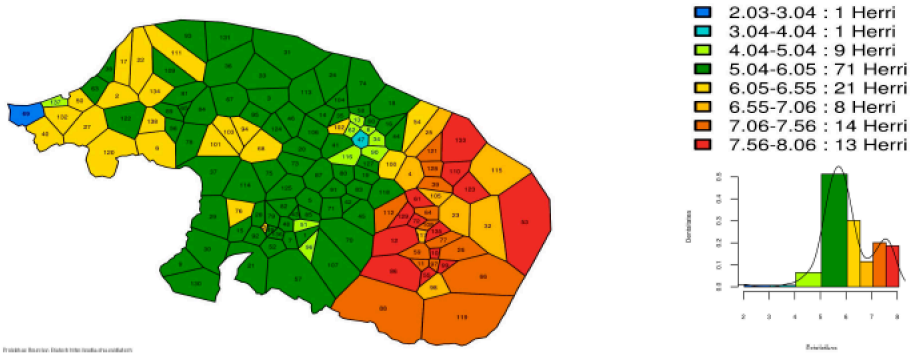
3.2.1.2 Transitional areas

We consider 'transitional area' an area whose *dialect* has been influenced by the *dialect* of one or more neighbouring local areas. Although many researches have been made about this subject, the concept has been used in different ways: sometimes small areas have been defined as transition areas, but in other cases large areas have been used.

DM, as a technique to measure linguistic distances, has also developed methods for analyzing transitional areas. These areas show a gradual change from one area to another. The standard deviation technique ('synopse des écarts-types' Fr.) is used in DM to detect transitional areas, located between two compact linguistic areas. It is known that the main feature of these areas is that they do not have many of their own features

©Universitat de Barcelona

and that they share the characteristics of their neighbouring localities and areas. The map 3 shows a large transition zone between two linguistic areas, situated one on the West and the other on the East (Map 3).



Map 3. Transitional zone in the Basque Northern area (Bourciez corpus)

3.2.1.3 Conservative areas

To find out where are the more linguistically conservative areas, quantitative dialectology uses the synopsis of skewness ('coefficient d'asymétrie de Fischer' Fr. / 'schiefe' Ger.). This statistical technique is used to analyze the symmetry and its orientation.

Let us assume that we are using a similarity distribution. In general, for the distribution of a certain location (linguistic identity) a positive asymmetry coefficient, a right skewed distribution, would indicate that most of the other locations have low similarity values. On the other hand, if the distribution happens to be negatively skewed, it would indicate that most of the other locations have a high similarity coefficient.

Note that if a dissimilarity distribution is used instead of a similarity one, the meaning of the skewness becomes the opposite.

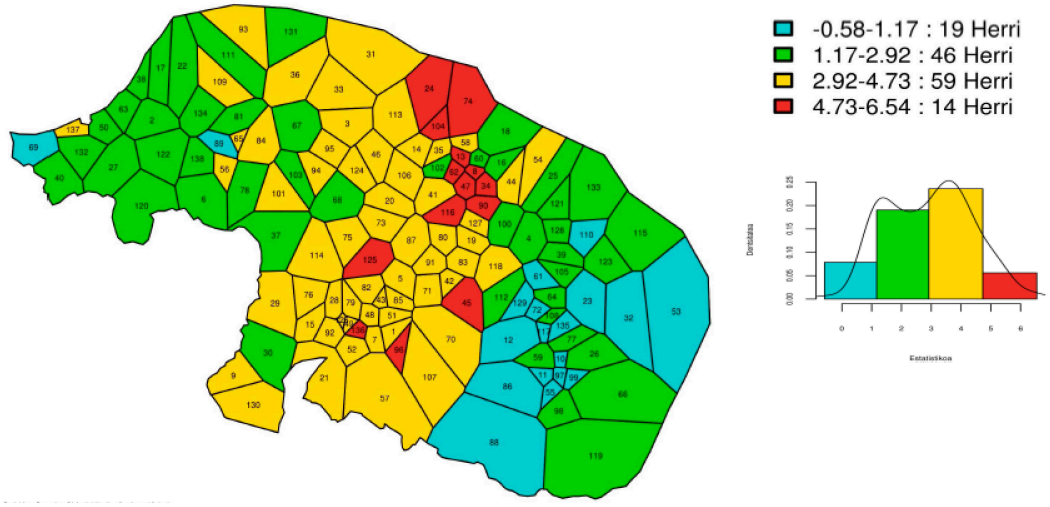
Taking into account the meaning of this technique two sides can be distinguished: one side has positive values (which fits with linguistically conservative zones) and the other one has negative values (which fits with non-conservative areas).

©Universitat de Barcelona

3.2.1.4

We define 'dialect core' as the group of localities that have the maxima of similarity distributions in the distance matrix of the data. These are the localities which have more similarities with the others. Quantitative dialectology has the 'Synopsis of the maxima of N similarity distributions' called technique to show the largest values of similarity of each locality. Its importance from the point of view of the dialectologist is the detection of the dialect core; that is to say, in these localities have the best relationship with the others in a bidirectional way, taking two localities in every distance. The higher similarity scores that a locality has with its neighbouring localities, the greater the dialectality of the locality is.

In Map 4 there are two groups of localities in red-orange colours: one on the Eastern part (red and orange colours) and the second one on the Western part (just one locality). These localities are extremely close from the linguistic point of view. Between these two groups there is a transitional area.



Map 4. Distribution of maxima of similarities (Bourciez corpus)

©Universitat de Barcelona

3.2.2 Isoglotic map (Honeycomb map)

Using the concept of quantitative isogloss (Goebl 1992), this procedure develops the traditional concept of isogloss and transforms it to include many features (not only one) and doing so in order to enable us to mark more or less important boundaries.

In the isoglotic map every side of the polygons of each locality turns into isogloss. Each side of the polygons takes the value of the linguistic differences between the localities of both sides. Thus, each isogloss has to be drawn with the corresponding characteristic.

3.2.3 Linguistic similarities (Beam map)

While the honeycomb map shows the linguistic differences between localities, the beam map shows the similarities. By drawing maps using this procedure the visualization of the linguistic proximities between localities it is shown and users can easily see the most linguistically homogeneous areas.

3.3 Multivariate DM

One of the most interesting achievements of the DM is the multivariate aggregative analysis. Among the large numbers of technical ways to carry out multivariate analysis in quantitative dialectology, cluster analysis and Multidimensional Scaling (MDS) techniques have been the most widely used.

3.3.1 Hierarchical classification of dialects (Cluster analysis)

There are many classes of clustering; from the Hierarchical, to flat clustering, hard and soft clustering (Prokic 2009), discrete clustering and composite clustering, Bootstrap clustering and Fuzzy clustering analysis, clustering with “noisy” (Dillon & Godstein 1984), etc.

©Universitat de Barcelona

However, whether we ask about the best clustering analysis probably the answer never will be a concrete and specific one. As Prokic (2009) said: "there is no one best clustering algorithm: every algorithm has its own bias […] The success depends on the data set it is used on […] Small differences in input can lead to substantial differences in output".

Diatech uses two types of cluster: the deterministic (hierarchical cluster) and the probabilistic fuzzy cluster. The first one determines the hierarchical classification of the localities, drawing clear cluster of localities, without taking into account the grade of the integration of each locality in the cluster. Meantime, the fuzzy cluster analyzes the grade of integration of each locality in the cluster; thus it draws better the linguistic boundaries of each cluster, showing whether the boundary is abrupt or on the contrary this boundary is not noticeable and more than a boundary is a transitional zone.

Respecting the algorithms used in the cluster analyses, three out of all the possible algorithms have been selected: Ward, Complete and Average, the three most frequently used ones in dialectometry. *Diatech* offers the possibility to choose the group length of the cluster, by selecting in the "group length" section the length which fits best according to the goal of the research. For example; 2 groups them 2 colours and 4 groups, them 4 colours.

Cluster analysis shows two outcomes: dendrogram and corresponding map. The dendrogram shows the hierarchy of the structure of data: grouping data in 6 groups, the green group is the most isolated by far. On the other hand, there is only one locality in dark blue (because of lack of data). It is not time to speak here neither about the best grouping of the dendrogram, nor the dialectal interpretation of these areas, but about that by selecting the group length we can change the configuration of the dialect areas.

3.3.2 MDS analysis

MDS allows us to visualize and simplify datasets with a large number of variables by reducing the dimensionality of the same ones with minimal distortion of the selected distance measure. This is really useful because it makes easier to represent the data in graphical terms.

©Universitat de Barcelona

The chosen particular method for MDS is known as Classical Multidimensional Scaling (MDS) or Principal Coordinates Analysis (PCA). This method is widely used due to the fact, that the loss function minimizes the selected linguistic distance (IRD, IPD, etc.).

The main advantage that PCA has is that observations can still be compared with simple graphical methods without the need of grouping (discretizing) the data. From a distance matrix (in this case 138x138), we are able to simplify another matrix with the selected dimensionality (K = 2, 138x2 in this case). A matrix that is 138 x 138 it is difficult to be expressed graphically; but one that is 138x2 it can be easily represented in a scatter plot.

It is also possible to represent it in a map by combining two basic colors such as green and blue and by adding a higher proportion of green for high values of the first dimension and similar for blue and the second dimension. This makes locations comparable with the need of grouping; therefore, the analyzed region looks 'continuous' rather than grouped. It represents the dialect continua, whereas cluster represents dialect areas.

Analyzing the map from left to right, for this result we can clearly see how there is some kind of linguistic continuum (smooth color changes) across the 'green' part of the map but how, suddenly, a jump occurs and a different area is drawn in blue. The same pattern can be seen in the scatter plot.

The Mardia measures the loss of the information when the matrix is changed from 138x138matrix to 138x2. The higher loss of information that happens, the lower Mardia coefficient that we will have; the "1" number indicates no loss of information, and "0" way too much loss.

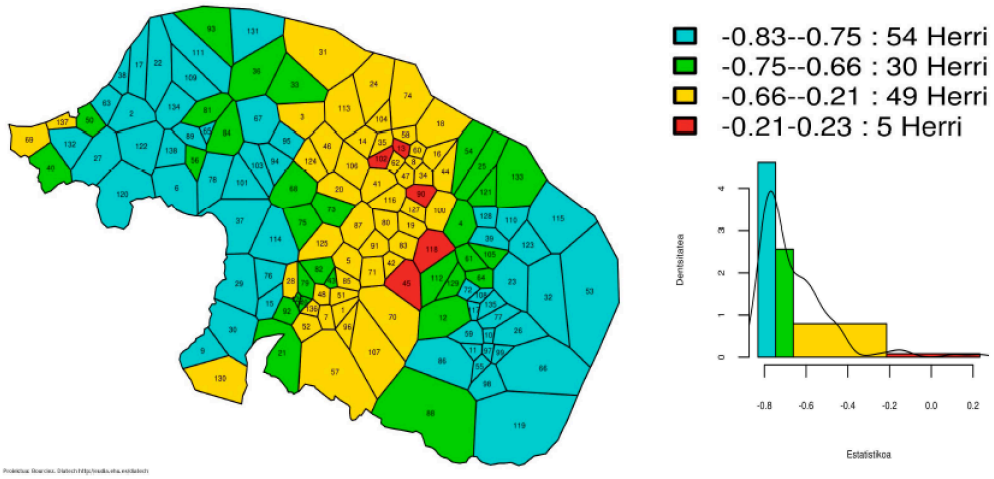
3.4 Correlative dialectometry

One of the most controversial subjects from the very first studies in dialectology has been the relation between the Euclidean distance and the linguistic differences. In effect, Séguy (1971) showed that the biggest Euclidean distance, the biggest linguistic difference. Nevertheless, this equation does not match perfectly in all situations and it

©Universitat de Barcelona

can change from one place to another, because of the humans' management of the space.

Goebl (2005) was the first dialectologist who implemented this kind of analysis into the VDM program, making it easy enough for his followers. *Diatech* has also implemented this technique and allows researchers to analyse the correlation between different dissimilarities. In this case (Map 5), left and right linguistic areas have better correlation with Euclidean distance than the central area has.



Map 5. Correlation between linguistic and Euclidean distance (Bourciez corpus)

"The VDM program also provides r(BP) values. If we look at these values next to the level panel, we see that all of them are positive. The coefficient r(BP) values generally ranges between -1 and +1. According to Aurrekoetxea (2010), whereas the first value indicates a negative correlation between compared variables (in this case lexical similarity and geographical distance), the second one signals a positive correlation. The "0" value indicates absence of any correlation.

3.5. Map, legend and histogram

All maps have a legend and a histogram; they have different information, depending on the map type (analytic or statistic one).

©Universitat de Barcelona

3.5.1 Legend

Legends are automatically created when a map such as an answer map, an analytic map or a statistic map is created. In the first case, legends are filled out with answers (orthographic, phonetic or lemmas) and in the second with intervals produced by visualization algorithms.

3.5.2 Histogram

There are different kinds of histograms. The histograms of *Diatech* show the relative density of each variable. The 'histogram' has three points to be explained: the length (or the height), the width and the line (or the curve). While the measurement of the similarity / dissimilarity (RIV, WIV, etc.) appears in the x-axis (axis of abscissas), in the y-axis (axis of ordinates) the relative frequency is shown, the density. That is to say, the high length (the height), for example, denotes that there are many localities in this section (Figure 2).

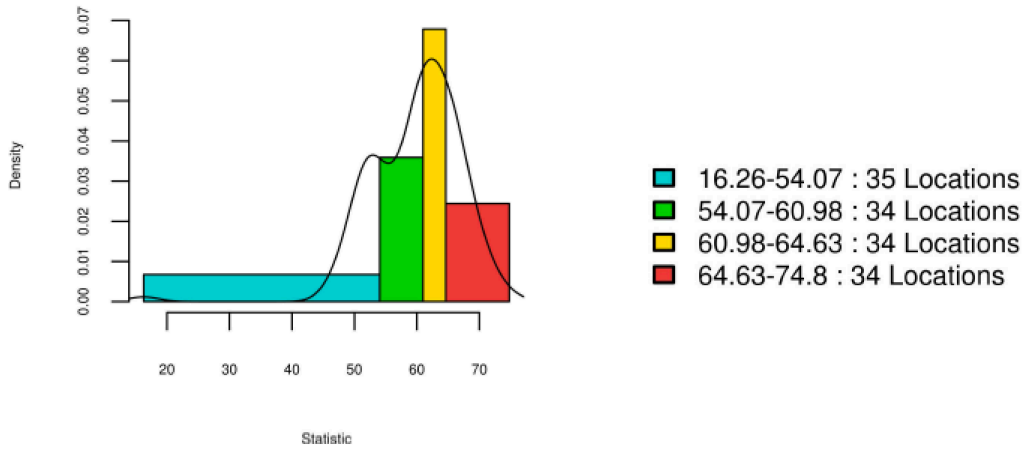


Figure 2. Histogram (Bourciez corpus)

In this case there are 35 localities in the blue area, but as the length between 16,26 and 54,07 is high (37,81), it is broad and it shows low frequency of localities in this group. The width of the green colour is 6,91 and there are 34 localities, so it is thinner but higher (the frequency of the localities is higher); the width of the yellow is 3,65 and

©Universitat de Barcelona

it also has 34 localities (the frequency of the localities is the highest); and, finally, the width of the red colour is 10,17 and there are 34 localities (the frequency of the localities is not very high).

And finally the line on the histogram, a Gaussian kernel estimation of the underlying distribution, indicates an estimation of the best choice of the numbers of groups: the best choice is when the curve and the height of each group agree or are very similar one to the other, by using the minimum of groups. It is a good indicator in order to choose the best grouping of the data.

4. Downloading data and outputs

On the one hand, *Diatech* provides the option to download the distance matrix (numerical distance between localities); and on the other hand, maps, legends and histograms. Underline, that all images have the jpg. format.

4.1 Distance matrix

The distance matrix is calculated in percentages; therefore, users can see the numbers and classify the distances from the smallest to the largest one. It is very important that users of DM come back to the linguistic data and explore which linguistic features have provoked the main distances among localities in different spaces. These distances will be different, according to the linguistic distance chosen, as you can see in the table.

4.2 Images

The *Diatech* tool provides two formats of images: images to be displayed on screens (RGB format) and the ones for editorials (CMYK format). On the other hand, users can also select the size of the images: medium or small. As it has just been mentioned, all images have the jpg. format.

©Universitat de Barcelona

5. Conclusions and future work

With the aim that DM should be available for everybody and socialize among them, *Diatech* puts at the disposal of dialectologists new technical improvements to make dialectometry easier and more comfortable, keeping in mind that well tracked techniques have to be used.

And about future works, the *Diatech* team reasserts that they will work hard to continue improving the tool in the future. We do not get out of our heads one of the linguistic motivations of the creation of *Diatech* tool: the transportability of the statistical outcomes, those we use in dialectology.

References

AURREKOETXEA, G. (2002) "Algunas consideraciones sobre la contrapregunta en las encuestas lingüísticas", in L. Rabassa (ed.), *Mélanges offerts à Jean-Louis Fossa,* Université de Toulouse II-Le Mirail, CerCLid 11/2, 57-65. <http://artxiker.ccsd.cnrs.fr/artxibo-00071400/en/>.

AURREKOETXEA, G. (2010) "The correlation between morphological, syntactic and phonological variation in the Basque language", in B. Heselwood & C. Upton (eds.), *Proceeding of Methods XIII,* Peter Lang, 207-118. <to see coloured maps: https://sites.google.com/site/eudiaehu/home/argitalpenak/argitalpenak-bourciez>

AURREKOETXEA, G., K. FERNÁNDEZ-AGUIRRE, J. RUBIO, B. RUIZ & J. SÁNCHEZ (2013) "DiaTech: A new tool for dialectology", *Literary and Linguistic Computing*, 28(1), 23-30.

DILLON, W. R. & M. GODSTEIN (1984) *Multivariate analysis–Methods and Applications,* New York: Wiley.

EUSKALTZAINDIA (2010-2016) *Euskararen Herri Hizkeren Atlasa,* vol. 1-6, Bilbao: Euskaltzaindia.

GOEBL, H. (1976) "La dialectométrie appliquée à l'ALF (Normandie)", in *Atti del XIV Congresso Internazionale di Linguistica e Dilologia Romanza* (Napoli 1974), Neapel, Amsterdam, vol. II, 165-195.

GOEBL, H. (1992) "Problèmes et méthodes de la dialectométrie actuelle (avec l'application à l'AIS)", in G. Aurrekoetxea & Ch. Videgain (eds.), *Nazioarteko dialektologia biltzarra.*

©Universitat de Barcelona

Agiriak / Proceeding of International Congress on Dialectology, Bilbao: Euskaltzaindia, 429-476.

Goebl, H. (2005) "La dialectométrie corrélative: un nouvel outil pour l'étude de l'aménagement dialectal de l'espace par l'homme", *Revue de Linguistique Romane*, 69, 321-355, + 24 maps.

GOEBL, H. (2007) "A bunch of Dialectometric Flowers: a Brief Introduction to Dialectometry", in Ute Smit, Stefan Dollinger, Julia Hüttner, Gunther Kaltenböck & Ursula Lutzky (eds.), *Tracing English through time. Explorations in language variation in honour of Herbert Schendl on the occasion of his 65th birthday,* Austrian Studies in English, v. 95, Wien: Braumüller, 133-171.

GOEBL, H. (2010) "Dialectometry and quantitative mapping", in Alfred Lameli, Roland Kehrein & Stean Rabanus (eds.), *Language and Space. An International Handbook of Linguistic Variation*, *Language Mapping*, vol. 2, Berlin: De Gruyter Mouton, 433-457.

GOEBL, H. (2013) "La dialectometrización del ALPI", in E. Casanova & C. Calvo (eds.), *Actas del XXVI CILFR Congreso Internacional de Lingüística y de Filología Románicas,* Berlin/Boston: Walter de Gruyter, 143-154.

GOOSSENS, J. (1977) *Inleiding tot de Nederlandse Dialectologie. Wolters-Noordho*, Groningen: Wolters-Noordhoff.

HEERINGA, W. (2004) *Measuring Dialect Pronunciation Differences using Levenshtein Distance,* in Groningen Dissertations in Linguistics, 46, <http://www.rug.nl/research/portal/files/9800656/thesis.pdf (last accessed 2014, August 4)>.

HEERINGA, W., J. NERBONNE, J. & M. R. SPRUIT (2007) "Associations among Linguistic Levels" <http://dare.uva.nl/document/98506> (last accessed 2014, August 4).

INOUE, F. (1996) "Computational Dialectology (1)", *Area and Culture Studies*, 52, 67-102.

KESSLER, B. (1995) "Computational dialectology in Irish Gaelic", in *Proceedings of the seventh conference of the European chapter of the Association for Computational Linguistics (EACL)*, San Francisco, CA: Morgan Kaufmann Publishers, 60-66. <arXiv:cmp-lg/9503002> (last accesed 2016, March 3).

NERBONNE, J., R. COLEN, Ch. GOOSKENS, P. KLEIWEG & Th. LEINONEN (2011) "Gabmap: a web application for dialectology", *Dialectologia*, special issue, II, 65-89. <http://www.publicacions.ub.edu/revistes/dialectologiaSP2011/>.

©Universitat de Barcelona

NERBONNE, J. & W. HEERINGA (2001) "Dialect areas and dialect continua" <http://www.let.rug.nl/heeringa/dialectology/papers/lvc01.pdf> (last accessed 2014-August 4).

PROKIC, J. (2009) "Clustering & Bootstrapping" <www.let.rug.n/~nerbonne/teach/rema-stats-meth-seminar/presentations/cluster-boot-prokic-2009.pdf> (last accessed 2014, August 4).

SÉGUY, J. (1971) "La relation entre la distance spatiale et la distance lexicale", *Revue de Linguistique Romane*, 35, 335-357.

Séguy, J. (1973) "La dialectométrie dans l'Atlas Linguistique de la Gascogne", *Revue de Linguistique Romane*, 37, 1-24.

©Universitat de Barcelona