Dialectologia. Special issue, VI *(2016), 191-221.*
ISSN: 2013-2247

Received 29 January 2016.

Accepted 23 March 2016.

ON QUANTITATIVE GEOLINGUISTICS: AN ILLUSTRATION FROM GALICIAN DIALECTOLOGY[1]

Francisco DUBERT & Xulio SOUSA

Instituto da Lingua Galega - Universidade de Santiago de Compostela

francisco.dubert@usc.es / xulio.sousa@usc.es

Abstract

The essential objective of dialectology, and especially geolinguistics, is the study of spatial linguistic variation, special relevance being given to the presentation of results through cartographic representations. The methods of geolinguistic data analysis focused for a long period on the description and evaluation of phenomena conventionally regarded as particularly relevant, either in isolated or in group cases, and the latter only when the discovery of the existence of limits between varieties was straightforward. The application of quantitative methods to geolinguistic studies began at the start of the 1970s and experienced a spectacular impulse over the last two decades. The development of new computer-based tools and the recognition that dialectal variation could not be reduced to simple characterizations encouraged researchers to employ new research methods, with which it was possible to understand in greater detail the patterns which function in geolinguistic variation. This present study seeks to provide an overview of the principal analytical techniques and representation of geolinguistic information developed during in recent years. Moreover, it intends to demonstrate, on the basis of some examples of application, the interest of these new methodologies and their usefulness for a better understanding of the spatial variety of languages.

Keywords

dialectology, geolinguistics, variation, statistical analysis, Galician

[1] This work was supported in part by the Xunta de Galicia and the European Union (under the grant GRC2013/40) and by the Ministerio de Economía y Competitividad (under projects FF12012-33845 and FFI2015-65208-P).

SOBRE XEOLINGÜÍSTICA CUANTITATIVA:

UNHA ILUSTRACIÓN PRÁCTICA DESDE A DIALECTOLOXÍA GALEGA

Resumo

A dialectoloxía, e de xeito especial a xeolingüística, ten como obxectivo esencial o estudo da variación lingüística espacial, dándolle unha especial relevancia á presentación de resultados por medio de representacións cartográficas. Os métodos de análise de datos xeolingüísticos centráronse durante moito tempo na descrición e análise de fenómenos considerados por convención especialmente relevantes, ben illadamente ou ben de xeito agrupado, e isto último só cando resultaba doado descubrir a existencia de límites entre variedades. A aplicación de métodos cuantitativos ó estudo xeolingüístico iniciouse a comezos dos anos setenta e coñeceu un espectacular impulso nas dúas últimas décadas. O desenvolvemento de novas ferramentas informáticas e o recoñecemento de que a variación dialectal non podía ser reducida a caracterizacións simples animou os investigadores á utilización de novos métodos de investigación, cos que foi posible coñecer máis en profundidade os padróns que funcionan na variación xeolingüística. Esta contribución pretende facer un repaso das principais técnicas de análise e representación de información xeolingüística desenvolvidas nos últimos anos. Ademais, búscase demostrar, a partir dalgúns exemplos de aplicación, o interese destas novas metodoloxías e a súa utilidade para un mellor coñecemento da variación espacial das linguas.

Palabras chave

dialectoloxía, xeolingüística, variación, análise estatística, galego

1. Introduction

The scientific study of the relationship between language and space goes back to the very beginning of dialectology as a discipline centered upon the analysis of linguistic variation. The connection between space and language was not only evident in the repercussion that sheer distance and physical barriers (mountains, rivers, lakes, deserts, etc.) impose upon the diffusion of linguistic variants, but also in the existence of different administrative, historical, cultural and social divisions which influence directly the expansion of linguistic features. Linguists soon realized that space should be observed and studied in the Euclidian dimension (the most evident) as much as in the social division, as well as that imagined by its speakers (Britain 2010). The interest of the observation of relationships between space and language already appears in the first studies on linguistic geography developed in Europe, whose interest is the enquiry

into the existence of similarities between ancient tribal divisions and dialect boundaries (Schrambke 2010). As a result, linguistic atlas projects which began to emerge from the mid-XIX century were not conceived so much as the result of geolinguistic research as basic instruments for undertaking an analysis of the relationship between the distribution of the features and linguistic varieties and highly heterogeneous extra linguistic factors, and also in order to assess the strength of proposals by the neogrammarians regarding linguistic change. Linguistic maps should serve to document the distribution in space of linguistic features and furthermore to help recognize the connections between linguistic information and other phenomena with spatial distribution.[2]

Geolinguistics, and in a general manner dialectology, focuses on the study of spatial linguistic variation, and affords special relevance to the presentation of results through cartographic representations. Maps are regarded as visualization tools which are extraordinary in their utility. European geolinguistic tradition was able to take special advantage of them in order to demonstrate the spatial distribution of lexical, morphological and phonetic variation. Researchers employed maps in order to seek explanations for phenomena of variation and linguistic change, to inquire into the history and spread of words and also to seek relationships between these linguistic phenomena and factors of another kind. In this research, maps were employed either in an isolated manner in order to document the distribution of sounds, words, etc., or as interpretative tools which helped discover areas which shared a set of dialectal features (*display maps* vs. *interpretative maps*, Chambers & Trudgill 1998: 25). This second function of maps helped in the identification of regional varieties which shared a series of features and also served to complement studies on the characterization and identification of linguistic domains. This task, fundamental in the development of dialectology during the XX century, was hampered by the very nature of linguistic factors, which only rarely display coincidental spatial distributions. In 1933, Bloomfield drew attention to this fact, which according to him was one of the fundamental

[2] It is necessary to distinguish between language maps, which account for the distribution of languages through territory, and linguistic maps, which display the spatial distribution of linguistic forms or features (lexical, syntactic, phonetic, etc.). This study will focus upon the latter type of linguistic maps.

obstacles which impeded the development of linguistic geography (Bloomfield 1933).

The analytical methods of geolinguistic data were linked for a long time to the simple description of phenomena regarded as especially relevant, whether in an isolated or group pattern, and the latter only when it proved straightforward to discover the existence of coincidences in spatial distributions. The identification of dialectal areas in a linguistic domain is based on the selection of a range of variables which allow singular and exclusive variants to be identified. In traditional dialectology, this procedure begins with the identification of a group whose features are reduced and which are considered linguistically relevant, and is completed by charting isoglosses upon a map. The limits between regional varieties should be determined by bundles of isoglosses, subsets of linguistic variables that display similar patterns of spatial variation. The proposals for dialectal divisions, and also for the identification of the limits between linguistic domains which are found in linguistic manuals of many of the Western languages, respond to this procedure. Commonly, the variables chosen for the identification and characterization of the regional varieties tend to belong to the fields of phonetics, morphology and to a lesser extent of lexicon and syntax. The problem presented by this method for identifying varieties, as Bloomfield has correctly pointed out, is that it proves to be extremely difficult to find linguistic variables which display similar spatial distributions. The alternative procedure which followed traditional dialectology was to select features which proved to be striking from the linguistic point of view; that is, variables that were especially significant: the dipthongation of Latin semivowels in peninsular Romances, consonantal mutations in dialects of German (*maken*/*machen* or *Apfel/Appel*), the distinction between /v/ and /b/ in European Portuguese, the behavior of unstressed /a/ and /o/ for varieties of Catalan or the formation of the plural in oxytones ending in *-n* which is the basis for the distinction between the three dialectal areas of Galician (Fernández Rei 1990).

The employment of a limited number of especially significant features is a principal applied repeatedly by traditional dialectology for the identification of varieties as much as for the establishment of limits between linguistic domains. The criterion for the identification of variables selected was not determined by the frequency of use of forms (morphemes, words, etc.) or of linguistic segments (sounds),

but rather by the characterizing distinctiveness attributed to them by scholars. A frank expression of this principle can be found in the clear affirmation of the Catalan grammarian and philologist Antoni Badia, when he seeks a differential criterion which allows him establish the dialectal limits between varieties of Catalan: "The critical point of the division of a domain in dialects is the establishment of a criterion according to which this division takes place: *it is best to adapt a single but very significant distinctive feature*" (Badia 1951: 70; emphasis added).[3] In spite of the fact that dialectologists resort to this simplification procedure for methodological reasons (they were not aware of tools for managing the large quantity of data provided by linguistic atlases), it is necessary to recognize that this resource could frequently be interpreted as a fallacy of incomplete evidence: the features for delimiting varieties are those which best serve to confirm and demonstrate a particular hypothesis, all the data which contradict this proposition are ignored (Wieling & Nerbonne 2015).

In recent decades a deep reflection began upon the methods of best employing materials obtained from projects of linguistic documentation and analysis of dialectal variation. Linguistic variation is a complex phenomenon, and therefore it proves inadequate to base analysis and description of the linguistic reality of a domain on a very reduced amount of data. Moreover, a procedure is necessary which would allow output to be taken from the large quantity of data on linguistic variation provided by linguistic atlases. The task of dialectology must go beyond individual analysis of certain data and the identification of features which it considers more significant or that are more adequate for its hypotheses. The researcher must be interested in discovering the existence of regularities, patterns of spatial distribution of variants, and for this an alternative method to traditional procedure is necessary, or at least a procedure which overcomes the difficulties in identifying bundles of isoglosses. During the last four decades, geolinguistics and dialectology were able to take advantage of the methods

[3] The same author recognizes that "although the most fundamental difference between Eastern and Western dialects is of the phonetic kind", but chooses another feature as one that is distinctive, given that according to him "it allows differential criteria to be unified, corroborates the aforementioned division between Eastern and Western Catalan, and confers more scientific value upon the dialectal division which previously had been practised by P. Barnils and that is rectified with the application of the new method" (Badia 1951: 70). It is evident that the selection is determined by a prior agreement regarding the dialectal division of the Catalan domain.

of statistical analysis and the visualization of data employed in other disciplines. It is true that linguistic variation is a complex phenomenon and affected by many variables, but it is equally true that it presents characteristics in common with other multidimensional realities and for which currently science has tools for especially beneficial analysis and description (Nerbonne 2006). Linguists are obliged to experiment with and apply these methods and also to assess if the results help them to understand the relationships between language and space, and in a general way linguistic variation, better.

2. Background and Development of Dialectrometric Studies

The methodological renovation in the field of geolinguistics took place fundamentally from the 1980s, with the development of computer applications which would allow the straightforward analysis of large bodies of data and to obtain representative graphic visualizations in a straightforward manner. The study of dialects by employing precise methods which combine computational and statistical approaches was given the name 'dialectometry', a term coined at the beginning of the 1970s by Jean Séguy (Goebl 2010b; Wieling & Nerbonne 2015). In spite of the effective birth of dialectometry taking place with the publication of Séguy's work on the *Atlas linguistique de la Gascogne* (ALG, Séguy 1973), the origin of this methodology is connected with the traditional procedure of searching for dialect boundaries on the basis of the identification of bundles of isoglosses. The Germanist Karl Haag was the first dialectologist who insisted that varieties employed for delimiting a linguistic region should be evaluated and assessed (Haag 1898). Furthermore, this research considered the need to use criteria to select the determining linguistic features of dialectal division: the number of words affected, the frequency in use of forms and the degree to which these change. Haag was also a pioneer in the detailed analysis of the value of isoglosses and in the employment of forms of innovatory representation (Goebl 2010a, 2010b; Schrambke 2010).

In the decades of the 1970s, the French dialectologist Jean Séguy coined and

initially developed the quantitative method of analysis of linguistic variation. The enormous variability that he came across when analyzing the materials of the *ALG* encouraged him to seek quantitative methods that would allow the interpretation of linguistic variation in a global manner. Séguy sought to transfer to dialectology the methods which had been employed for some time in other scientific disciplines such as biology, economics, population genetics and psychology. These techniques enabled him to measure the linguistic distance between lects and to determine dialect boundaries on the basis of localities which presented the highest linguistic distance values. The results of his studies caused him to re-assert the idea that linguistic domains are not mosaics of dialects, but continual spaces in which linguistic changes gradually accumulate. The premature death of Séguy prevented his conjectures from becoming a mature and well-defined method.

The consolidation of dialectometry as a method and discipline of dialectal studies took place because of the work undertaken by Hans Goebl. This Austrian researcher began work on classification and numerical analysis of dialectal data at almost the same time as Séguy (Goebl 1971, 1975, 1976). From his first studies, Goebl combined procedures of numerical classification belonging to statistics with genuinely innovatory visualization techniques. His specialization as a Romanist meant that his studies focused initially on the Romance field, a linguistic domain which he regards as unique and therefore which must be studied in a different manner (Goebl 2010b). In his work he applied dialectrometric methodology to different Romance areas: Gallo-Romance (Goebl 2000, 2002, 2003), Italo-Romance (Goebl 2008) and more recently Ibero-Romance (Goebl 2013a, 2013b) also. Moreover, he also experimented in the comparative analysis of dialectal, onomastic and genetic data (Goebl et al. 2005). Hans Goebl founded a dialectrometric school in his research centre in Salzburg, specializing in the study of Romance linguistic space. In the setting of this school he developed a computer program *Visual Dialectometry*, which allows a large number of statistical calculations to be made and to obtain graphic visualizations in a straightforward manner (histograms, dendrograms, choropleth maps, honeycomb maps, beam maps, etc.). The Salzburg dialectrometric school is based on the application of quantitative (statistics) and cartographical (visualizations) methods to data of any linguistic atlas

seen categorically (Goebl 2006 and 2010b). The fundamental contributions of the school are the employment of frequency weightings in order to calculate aggregate differences, the introduction of a variety of descriptive statistics as well as cluster analysis, and the use of Thiessen polygons as a means of dividing maps into areas related to the localities surveyed. The work undertaken over the last three decades by Hans Goebl was an essential stimulus for dialectometry to develop and spread as a fundamental method in geolinguistic studies.

At the beginning of the 21st century, John Nerbonne and Wilbert Heeringa began to publish a series of studies which represented the constitution of what ended up being referred to as the Groningen School of Dialectometry. Following the route began by Goebl, they experimented and developed new analytical techniques based on different statistical procedures and in addition incorporated the numerical measure of pronunciation distance (Nerbonne et al. 1999). The statistical methods in which the research of this school focused are those of multidimensional scaling, hierarchical agglomerative clustering and principal component analysis. The studies by authors connected with this school contributed to dialectometry attaining its theoretical bases and to it broadening the areas and perspectives of application (for a description of the principles and methods of the two schools, consult Goebl 2010b, Szmrecsanyi 2014 and Wieling & Nerbonne 2015).

During these years, the development of dialectometry also enjoyed contributions from outside the Germanic area. In addition to studies on the application of techniques proposed by the Salzburg and Groningen schools in different linguistic domains, the contributions by Kretzschmar in the United States, Kirk and Thomas in the United Kingdom, Cichocki in Canada and Inoue, Kassai and Ueda in Japan are to be highlighted. In recent years, new theoretical and technical contributions, which assert the strengthening of the discipline in the field of dialectal studies, continue to be added. However, it must be recognized that the application of quantitative methods in geolinguistic research still continues to come across sufficient objections and skepticism today within linguistics. Goebl points out that this distrust towards dialectometry comes from ignorance, from the interpretation as an attack upon qualitative treatment of the data and also from the consequences of the eternal

conflict between the designated "human" sciences and "pure" or "natural" sciences (Goebl 2003: 89-90).

3. Tools

The theoretical development of dialectometry ran parallel to the incorporation of computer tools for the analysis of geolinguistic information. The most widespread and employed programs of statistical analysis can be applied with few adaptations to the study of linguistic variation. The open-source statistical package R (http://www.R-project.org), widely used among statisticians and data miners, is commonly used in order to obtain figures which imply measures of distance, similarity, cluster analysis and many of the more sophisticated analyses. In spite of it being a program which requires training and some experience, the investment is worthwhile, given that many researchers openly share code, script and data used in their studies. R packages such as *rMaps* make it easy to create, customize and share interactive maps from R (http://rmaps.github.io/). Research in dialectometry also employs many data visualization techniques which represent core methods of GIScience, a scientific discipline which has undergone extraordinary developments in recent decades, and which signifies a popularization of its methods and also an extending of the use of GIS tools. GIS software systems currently offer many applications which can be employed in a straightforward manner for the representation of geolinguistic information or rather the results obtained from the statistical analysis of these materials (Sibler et al. 2012; Chun & Griffith 2013). In addition to these widely used tools in different scientific fields, at the heart of some projects on the quantitative study of linguistic data, specific applications were designed and developed for dialectrometric research. Presently a brief review of the most common and employed tools currently in the scientific community will be made.

3.1 VisualDialectometry (VDM; http://ald.sbg.ac.at/dm/germ/VDM/)

Parallel to the beginning of dialectrometric analyses of different Romance linguistic atlases (1997-2000), Hans Goebl and Edgar Haimerl designed special purpose-built software called *VisualDialetometry* (VDM). In order to use this program, the information extracted from a given linguistic atlas (map names, linguistic data, places, taxats, map limits, etc.) must be incorporated in the form of Access tables. The linguistic variants associated with each one of the maps are treated as categorical data and have to be classified (typized or taxated via "taxation") according to traditional levels of linguistic analysis (phonetics, vocabulary, syntax, morphology, etc.). The basis for the cartographic representation of the results, as in the rest of the dialectrometric tools, is a polygon map (Voronoi tessellation) in which each polygon functions as a dialectal cell corresponding to each one of the enquiry points of the linguistic atlas. Through a very friendly interface, data can be analyzed with statistical techniques popularized by the Salzburg school and the respective graphic and cartographic visualizations obtained: working maps (thematic maps), similarity maps, parameter maps, interpoint maps, beam maps, cluster analysis, correlative analysis, etc. (Goebl 2010b). The results can be exported as images and data tables. The VDM was for a period one of the tools most used for the dialectrometric analysis of different linguistic domains, especially Romance. The tool functions only with the MS Windows operating system and expert assistance in programming is required for it to be adapted to a specific linguistic domain.

3.2 Gabmap (http://www.gabmap.nl)

A web application, which began as an online user-friendly version of RUG/L04[4] (Nerbonne et al. 2011; Snoek 2014), was developed at the core of the Groningen School of Dialectometry. *Gabmap* can work with categorical data (phonetical, lexical, morphological, etc.) or numerical data (formant frequencies, etc.). Data can be inputted into the program as a tab separated text file, as the cartographic bases (maps) must be inputted in kml format (XML notation for expressing geographic

[4] RUG/L04, developed at the University of Groningen by Peter Kleiweg, was employed for years as a standard computational application in dialectometry. Its initial use was limited to determining the geographical distribution of aggregate differences in dialectal phonetics (on the basis of edit distance).

features), which facilitates use by researchers without specific knowledge of programming. It also contains data-inspection tools which help to find errors in the data and allow the geographical distribution of single features to be visualized, although it is not possible to create a thematic map of variants (the working map in VDM). *Gabmap* has allowed users to investigate aggregate dialect patterns using cluster analysis and MDS; in addition, it allows users to detect characteristic features of dialect regions. The results of the calculations and the graphics can be downloaded easily. *Gabmap* offers complementary tools to a VDM, but the opportunities for the exploitation of data and of visualization are less. The user-friendly nature of *Gabmap* is turning it into one of the most employed programs today for analysis of dialectal data from very diverse sources.

3.3 DiaTech (http://eudia.ehu. es/diatech)

DiaTech is another web application created by a multidisciplinary team of UPV/EHU, led by Gotzon Aurrekoetxea, an expert researcher in Basque dialects with experience in the use of VDM. *DiaTech* follows the Salzburg school´s analytical model and focuses on the types of analysis present in VDM. In terms of how it works for the creation of new projects and the data load displays many similarities with *Gabmap* (Aurrekoetxea et al. 2013). Unlike other tools, it enables the analysis of multiple responses, namely multiple variants attributed to single sites or different respondents. The application has a friendly work environment which allows uploaded data (data files and locations) to be consulted and edited with ease. Data can be analyzed by using different statistical and visualization procedures: synoptic maps, beam maps and cluster analysis. The results of the projects realized with *DiaTech* can be downloaded or even made public from the same page of the application. The project to which the application is linked is still in development, therefore some of the user tutorials and help facilities are incomplete.

3.4 GeoLing (https://www.uni-ulm.de/en/mawi/geoling)

GeoLing is a manageable tool for performing statistical analysis on spatial-dialect data. The program was created within the *New Dialectometry Using Methods of Stochastic Image Analysis* (Universität Ulm, Universität Ausburg and Universität Salzburg) project and was originally designed to be used with data from the *Sprachatlas von Bayerisch-Schwaben*. The software is written in Java and runs on multiple platforms (Windows, Linux and Mac) and enables analyses based on techniques from geostatistics, image analysis and data mining (factor analysis, density estimation, cluster analysis, etc.). The application runs on the basis of data in the form of a SQL database or rather with comma-separated values files. Unlike other dialectrometrical software, *GeoLing* allows the study of the relationship between individual variants and whole varieties for the purpose of recognizing recurring patterns in variants distribution and identifying hidden structures and large-scale patterns in the aggregate data. Moreover, the program helps to detect groups of maps that share spatial features. Of the range of dialectometry applications mentioned, it is the most recent and least used. However, and in spite of its use requiring certain training, the results obtained in its application are genuinely interesting (Pickl & Rumpf 2012, Pickl 2013, Pickl et al. 2014).

4. Studies in Dialectometry in Galician

The application of quantitative techniques to studies of Galician dialectal varieties began in 2001 with the presentation of the work by João Saramago on the lexical differentiation in Galician and Portuguese linguistic domains (Saramago 2002). Forty-five lexical variables were analyzed by using the methods of the Salzburg school and on the basis of data extracted from the Portuguese materials from the *Atlas Lingüístico de la Península Ibérica* (ALPI) and the Galician material from the *Atlas Lingüístico Galego* (ALGa). The author uses the calculation of the distance values and the visualization of results to discover the coincidence between political divisions and

dialectal limits and also to recognize the lexical affinity existing between Galician varieties and those of the northwest of Portugal (Saramago 2002: 64). In 2003, another study on dialectrometric analysis in the Galician linguistic domain was presented, also on the basis of the *ALGa* materials, in which a first proposal for the dialectal division of varieties was made by employing the cluster analysis procedure (Sousa 2006). This work is the first result from a project on dialectrometric analysis of varieties of Galician carried out at the Instituto da Lingua Galega from 2004 [Galician Language Institute].[5] To follow, some procedures for the dialectrometric method on the basis of studies related to this project and others are set out. Calculations and visualizations commented on were carried out with R and VDM.

4.1 Analysis of similarities and interpuntal differences

The statistical treatment of large bodies of non-hierarchical data enables the performing of different analyses, impossible in traditional qualitative dialectology, which is forced to work with a cognitively-limited number of features, prioritized by the researcher in a subjective and conventional way.

For the greater part of the SD (Salzburg Dialectometry) applications, the basic methodological features are the designated interpuntal similarity indexes. Once the taxation study is performed and transformed into numerical indexes, the qualitative data of the work maps which are used, a comparison is made of the responses obtained for each question at the different locations of the network with responses to the same question at the other locations.[6] For example, after comparing the responses at point C.37 of the *ALGa* with the other points in the database of Dubert (2011), it was confirmed that C.37 obtains a similarity index of 91.74% with C.34, of 90% with C.22 and C.43, or of 89.57 with C.27, and so forth, until C.37 is compared with another 166

[5] On the project´s webpage (*Análise dialectomética do datos do Atlas Lingüítico Galego*, <ilg.usc.gal>), general information can be found regarding the published results.

[6] The SD method and the VDM program only allow for one response per point, so that in the case of having more than one response per point, the researchers who design the database must select, according to some criterion, just one of the responses gathered at the point. It should also be noted that at one point no response was obtained, the lack of response interpreted as a differet response. Therefore, work maps with many points without response should be avoided wherever possible.

remaining points.

These indexes now allow the development of maps like those of Figures 1 and 2, denominated *similarity profiles*, which show, respectively, the distribution of points similarity indexes L.28(77) and L.31(80), obtained from the database of Dubert (2011). What can be seen on the map are groupings of polygons (each one of them representing a point on the network) and grouped by a colour for its degree of similarity with the reference point. The maps are accompanied by a histogram which on the abscissa provides information regarding the range of similarity indexes and from where the segmentations of this continuum take place, and on the ordinates line provides information regarding the number of points contained in each segment. Accompanying it, there is also the map of a legend based on the histogram and on which the range of similarity indexes associated with each colour is indicated.

It can be noted that these maps display a grade of similarities between a point of reference and all other points of the network. As can be appreciated in the histograms, each segment recovers a range of similarities; in this case, the bright colours for the points with the highest similarity indexes and the dull colours for the lowest indexes. It is a mathematical algorithm provided by the VDM program, which divides the continuum, shapes the segments and attributes a colour to them. However, nothing in the theory tells us into how many segments the continuum of the indexes must be divided; the quantity of segments depends on the researcher´s requirements and the perceptive ability of the human being.

Furthermore, on these two maps, another clear illustration of the geolect can be seen: L.28 (77) and L.31 (80) are two contiguous points which share a very high similarity index; however, the red area, of high similarity indexes of L.28 (77) looks northwards; that of L.31 (80) southwards.

This same method allowed Sousa (forthcoming) to measure the linguistic distance between the standard variety (whose data, taken from ILG/RAG 2003, was included in a fictitious geographical point Ga1) and the 167 geographical varieties analyzed in the *ALGa*. Figure 3, taken from Sousa (forthcoming), was carried out with 324 morphological work maps. In this case, as the histogram shows, the red and green colours are allocated to the points with dialects closest to the standard variety, and the

blue and violet colours to the most removed points. It should be noted, however, that blue colours accumulate similarity values that range from 71.21% to 74.77%. As can be seen, 132 points out of 167 have similarity indexes greater than this 71.21%.

With a similar method, Fernández Rei, Moutinho & Coimbra (2014), produced the map contained in Figure 4. This map allows a comparison of the intonation collected at a Portuguese point of their database on the intonation on total interrogative sentences with another 23 Galician and Portuguese points. This map displays that the Portuguese point 003 Moledo of their network "presents a small distance with the Portuguese varieties and also regarding the Galician ones from the Rías Baixas. On the other hand, the distance with relation to other Galician varieties is quite high" (Fernández Rei, Moutinho & Coimbra 2014: 129).

The *VDM* application enables the production of maps such as that of Figure 5, designated as *honeycomb maps*. On these maps, the result of the statistical analysis of the differences between contiguous points is represented. After processing its computations, *VDM* offers another continuum of values of interpunctual differences, as can be observed in the histogram, where also the divisions of the continuum appear together with the colour which is given to each segment. The lines of the map, in accordance with the segmentations of the histogram, denote the degree of difference which exists between two contiguous points. Once again, the colour of the lines depends on the quantity of difference: dull colours when they present high values of difference and bright colours when they present low values. Figure 5 shows an isogloss map[7] of these characteristics, made with 121 morphological maps (Álvarez, Dubert-García & Sousa 2006). As can be seen, the north of Galicia and the central-western zone are areas of great homogeneity, as the interpunctual differences are low (the red lines connect points with difference values which oscillate between 2.7% and 10.91%). On the other hand, it can be seen that the southwest and the east display a large internal fragmentation (with ranges of difference values between 20.01% and 51.52%).

[7] Although these maps have a certain formal similarity with an isogloss map, this does not mean that the lines are similar to the isoglosses. Isoglosses are lines "drawn on a dialect map to represent a fairly sharp boundary between two competing linguistic forms used in neighbouring areas" (Trask 2000: s.v.). On the honeycomb maps the lines refer to statistical differences obtained by computing many (at times, hundreds) of different limguistic phenomena.

An important element of the method is that the similarity profiles as well as the honeycomb maps are produced on the basis of an aggregate data analysis, in which all variables have the same weight. Clearly, this fact means that subjectivity in selecting this or that isogloss in order to establish areas can be avoided.

4.2 Synoptic maps of average and standard deviation

On previous maps, similarity and difference values of one point against others were compared. However, these similarity indexes can be reworked and employed in different ways through maps referred to as *synoptic*, which show structures underlying the entire points network. For example, it is possible to obtain the average value of all similarity indexes for each point. Goebl does not tend to comment on these maps, but it is felt that they contain important information[8] and that they allow for a better interpretation of the maps which represent typical deviation indexes of these averages. In fact, these maps are presented together in Figures 6 and 7. Figure 6, which charts the similarity averages of each point, displays a central zone in Galician with high averages, which indicates that these points share high similarity indexes with other points; on the other hand, the periphery, i.e., the north-western, south-western and eastern dialects show low averages, which reveals that its similarity indexes are low.

For its part, Figure 7 shows the map and graphs which correspond to the standard deviation values of the similarity averages contained in Figure 6. On revising this map in Figure 7, it can be confirmed that the centre of Galicia has more high averages of low typical averages, which display a concentration of high similarity values around the high averages: these points share many similarities with others. On the other hand, in the north-western dialects as well as in Asturias, the averages are low and the typical deviations high, as the values are more dispersed around the average. These are areas with much dialectal character, as they contain groups of

[8] "Dans une perspective communicative [...], la moyenne arithmétique d'une distribution de similarité peut être utilisée à en évaluer numériquement la position central au sein du réseau examiné" (Goebl 1981: 389).

points which resemble each other considerable, but that are very different to others. Finally, in the southwest as well as in the southeast there are low averages and typical low deviations: this is the result of a concentration of low similarity values around averages of low similarity, which implies that all these points are in general quite different to the rest (both those near and removed).

The usual dialectrometric interpretation of this great central space with greater similarity indexes is that this zone constitutes an area of transition located between areas with more compact and divergent features:

> A définir ce qu'est une zone de transition ou un parler-pont, nous dirions qu'il s'agit des parlers intermédiaires situés entre les noyaux dialectaux signalés ci-dessus. La caractéristique de ces parlers est de posséder peu de caractères propres et d'offrir peu à peu le 'chemin' ou 'pont' qui permet de passer d'un système linguistique à un autre (Aurrekoetxea & Videgain 2009: 101).

4.3 Correlative dialectometry

Quantitative analysis techniques allow the relations that exist between two data distributions to be discovered. The method employed in this case, referred to as correlative dialectometry, enables the visualization and comparison, for example, of the geolinguistic relationships between the distribution of phonetic data and another of morphological data; and it also enables an analysis of the relationship between linguistic and geographical distances. In this way, it can be ascertained as to whether all dialects maintain the same quantitative relationship of similarity between geographical variations of two language components; or between any linguistic component and geographical space.

The map of Figure 8 is the result of a comparison of C.37 with the rest of the points of the *ALGa* (Dubert 2011). As can be seen, as C.37 is approached geographically, the linguistic similarity indexes increase. On the Map of Figure 9, the other points of Galicia are grouped together according to their geographical proximity in regard to C.37. As can be appreciated when comparing the two maps, the increase of linguistic similarity indexes takes place more or less in tandem with the increase of geographical proximity indexes. On the other hand, if the maps of Figures 10 and 11

are examined, it can be seen that with a comparison of O.10 (98) with the other points of the *ALGa*, such a correspondence between the geographical proximity indexes and the linguistic similarity indexes does not exist. Linguistic distribution shows that the high similarity indexes extend only towards the north and west, where a somewhat abrupt fall occurs, in spite of the geographical proximity between O.10 (98) and the contiguous points to its left.

VDM software makes it possible to calculate, for each point of the network, the co-efficient of the correlation of Bravais-Pearson; the object of this co-efficient is precisely to measure, for each point, the degree to which its variations in a distribution (linguistic similarity) are accompanied by similar quantitative variations in the other (geographical proximity). In order to obtain this co-efficient, *VDM* analyzes the degree of relationship which exists, at each point, between its linguistic similarity indexes and its geographical proximity indexes. Once the co-efficient of all points have been obtained, it is possible to show their geographical distribution on a map.

The correlation index values range from 1 to -1. The value 1 indicates a direct correlation: the highest values of X (geographical proximity), and values higher than Y (linguistic similarity); the value 0 indicates autonomy between X and Y; the value -1 indicates an inverse correlation: the highest values for X, the lowest values for Y.

The analysis produces another continuum of 167 correlation indexes, which can be divided automatically into various segments through a mathematic algorithm. Once again, the bright colours are associated with the segments with indexes higher than 1, and the dull colours are applied to segments with indexes further away than 1. In this case, the bright colours represent indexes closest to the direct correlation (the most X, the most Y); the dull colours, indexes more distant from the direct correlation. Figure 12 shows the distribution in the space of the correlation indexes of Bravais-Pearson amongst the linguistic similarities with phonetic data and the geographical proximity obtained for each one of the points of the *ALGa* (Dubert 2011, 2012). As can be seen, a quite symmetrical distribution of bright and dull colours can be seen on the map: the centre of Galicia is shown as a place where the indexes most removed from the direct correlation appear, in such a way that on them a greater autonomy between geographical proximity and linguistic proximity can be verified. As Goebl points out, in

the areas painted in dull colors

> l'étalement des similarités linguistiques dans l'espace a été libéré complètement des contraines euclidiennes de l'espace, évidemment por des raisons sociales et politiques de toute sorte (Goebl 2008: 101).

In any event, it is not known to what extent one can assert that in the case of Galician all points with dull colours display a great autonomy between geolectal variation and geographical distances, as the green colours begin to appear with a value of 0.64 for the Bravais-Pearson index. This shows that in these areas there also exists an appreciable tendency towards positive correlation between linguistic similarity and geographical proximity. Nonetheless, northwestern and eastern Galician have the highest correlations between geographical distance and linguistic difference Dubert-García (2013) provides a similar analysis for the correlation between phonetic and morphological similarities.

4.4 Cluster analysis and dendrograms

Another technique, more in keeping with traditional dialectology methods, but impossible to carry out without work on variables quantification, is cluster analysis. In order to perform cluster analysis, *VDM* begins by pairing off the most similar localities; having created these pairs, it then proceeds to group the pairs with others that are most similar to them, and so on until all localities have been connected. This analysis creates a binary hierarchy of groupings which is represented on the dendrograms. The closer that a fork of the main tree trunk is, the more heterogeneous the grouped points are. The greater the distance from the trunk, the more homogenous the points are. The map is the result of the cartographic representation of clusters which are observed in the dendrograms.

Figure 13 shows a cluster analysis carried out concerning data from the *ALGa* with 189 work maps, mostly phonetic and morphological (Sousa 2006). Although on the map three groupings of polygons similar to those of the territorial divisions of Galician in three blocks distributed from west to east can be seen (Fernández Rei

1990), the map cannot be distributed in the same way. On the bidimensional surface of the map appears a hierarchical structure which is obtained by comparing the distributions of colours of the map with the dendrogram clusters. With this being a binary structure, the first large division takes place between western Galicia and the remainder of the country; this remainder is divided, in turn, into a Central Galicia which is hierarchically closer to Eastern Galicia; these two face as a block Western Galicia, which proved to be singled out in the first partition.

The number of clusters and subdivisions is an open question. Two groupings can be made (which would divide the Galician-speaking territory into two, western and the remainder). In principle, no degree of subdivision is theoretically relevant when identifying dialectal groups, as an ideal number of clusters does not exist which shows in a clear manner *which are the dialects of Galician*. However, whilst compact groups appear, the method does allow an identification of what areas share amongst themselves a significant number of linguistic similarities (areas established on the basis of a higher set of data and not of some isoglosses) and, therefore, it is an interesting complement to know the geographical distribution of groupings of dialects.

Fernández Rei, Moutinho & Coimbra (2014) present their work in an analysis of similar groupings regarding intonation of Galician and Portuguese points. As can be appreciated in their dendrograms (p. 129), some Galician points from the Rías Baixas (om1 [O Grove], on1 [Cangas] and oo1 [Oia]) are closer, within the binary structure, to the Portuguese points than the remaining Galician points (Fernández Rei, Moutinho & Coimbra 2014: 129).

5. Conclusions and remarks

In spite of the limitations and all the precautions that must be taken when accepting its results (see Dubert-García 2011), in the authors´ opinion, dialectometry surpasses some of the operational restrictions of traditional dialectology. Firstly, it allows researchers to work in an objective manner (i.e., without subjectively prioritizing some linguistic features over others) with a large body of data (i.e., without

having to just accept the tracing of some isoglosses of some isolated linguistic features). Therefore, it allows for an exhaustive application of information contained in the linguistic atlases. Secondly, the statistical analysis of large bodies of data shows hidden patterns and structures, immanent and hidden in the body of data, which would be impossible to recover or perceive in qualitative dialectal studies. This does not prevent the employment other social, geographical, historical etc. information from being necessary for the complete and coherent interpretation and explanation of these structures.

References

ÁLVAREZ, R., F. DUBERT-GARCÍA & X. SOUSA FERNÁNDEZ (2006) "Aplicación da análise dialectométrica aos datos do *Atlas Lingüístico Galego*", in R. Álvarez, F. Dubert-García & X. Sousa Fernández (eds.), *Lingua e territorio*, Santiago de Compostela: ILG/CCG, 461-493.

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

AURREKOETXEA, G. & C. VIDEGAIN (2009) "Le project Bourciez: traitement géolinguistique d'un corpus dialectal de 1895", *Dialectologia*, 2, 81-111. <http://www.publicacions.ub.edu/revistes/dialectologia2/>

BADIA i MARGARIT, A. (1951) *Gramática Histórica Catalana,* Barcelona: Noguer.

BLOOMFIELD, L. (1933) *Language,* New York: Holt.

BRITAIN, D. (2010) "Conceptualisations of geographic space in linguistics", in A. Lameli, R. Kehrein & S. Rabanus (eds.), *Language and Space. An International Handbook of Linguistic Variation. Volume 2: Language Mapping*, Berlin: de Gruyter, 69-102.

CHAMBERS, J.K. & P. TRUDGILL (1998) *Dialectology*, Cambridge, UK: Cambridge Univ. Press, 2nd edition.

CHUN, Y. & D.A. GRIFFITH (2013) *Spatial Statistics and Geostatistics: Theory and Applications for Geographic Information Science and Technology*, Thousand Oaks, CA: Sage.

DUBERT-GARCÍA, F. (2011) "Developing a database for dialectrometric studies: the ALGa phonetic data. Dialectrometrical analysis of 230 working maps", *Dialectologia et Geolinguistica*, 19, 23-61. DOI: 10.1515/dig.2011.002

DUBERT-GARCÍA, F. (2012) "Geographic proximity and Linguistic similarity. A correlative analysis of Galician phonetic data", in X.A. Álvarez Pérez; E. Carrilho & C. Magro (eds.), *Proceedings of the International Symposium on Limits and Areas in Dialectology (LimiAr), Lisbon 2011*. Lisboa: Centro de Linguística da Universidade de Lisboa. http://limiar.clul.ul.pt, 71-91.

DUBERT-GARCÍA, F. (2013) "An analysis of Galician Dialects in Correlative Dialectometry", in X.A. Álvarez Pérez, E. Carrilho & C. Magro (eds.), *Current Approaches to Limits and Areas in Dialectology,* Newcastle upon Tyne: Cambridge Scholars Publishing, 171-198.

FERNÁNDEZ REI, E., L. DE CASTRO MOUTINHO & R. L. COIMBRA (2014) "As entoacións galega e portuguesa: a fronteira á luz da dialectometría e da percepción", in X. Sousa, M. Negro Romero & R. Álvarez (eds.), *Lingua e identidade na fronteira galego-portuguesa*, Santiago de Compostela: CCG, 115-141.

FERNÁNDEZ REI, F. (1990) *Dialectoloxía galega,* Vigo: Xerais.

GOEBL, H. (1971) "Projekt einer sprachstatistischen Auswertung von in Sprachatlanten gespeicherter linguistischer Information mit Hilfe elektronischer Rechenanlagen", *Linguistische Berichte*, 14, 60-61.

GOEBL, H. (1975) "Dialektometrie", *Grazer linguistische Studien*, 1, 32-38.

GOEBL, H. (1976) "La dialectométrie appliquée à l'ALF (Normandie)", in A. Várvaro (ed.), *Atti del XIV Congresso Internazionale di Linguistica e Filologia Romanza*, Nápoles: Macchiaroli, Ámsterdam: Benjamins, vol. 2, 165-195.

GOEBL, H. (1981) "Éléments d'analyse dialectométrique (avec application à l'AIS)", *Revue de Linguistique Romane*, 45, 349-420.

GOEBL, H. (2000) "La dialectométrisation de l'ALF: présentation des premiers résultats", *Linguistica*, 40, 209-236.

GOEBL, H. (2002) "Analyse dialectométrique des structures de profondeur de l'ALF", *Revue de Linguistique Romane*, 66, 5-63.

GOEBL, H. (2003) "Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur", *Estudis Romànics*, XXV, 59-121.

GOEBL, H. (2006) "Recent advances in Salzburg dialectometry", *Literary and Linguistic Computing*, 21, 411–35.

GOEBL, H. (2008) "Brève introduction aux problèmes et méthodes de la dialectométrie", *Revue roumaine de linguistique*, LIII, 1-2, 87-106.

GOEBL, H. (2010a) "Introducción a los problemas y métodos según los principios de la escuela dialectométrica de Salzburgo (con ejemplos sacados del "Atlante italo-svizzero")", in G. Aurrekoetxea & J.L. Ormaetxea (eds.), *Tools for Linguistic Variation*, Bilbo/Bilbao: UPV/EHU, 3-39.

GOEBL, H. (2010b) "Dialectometry and quantitative mapping", in A. Lameli, R. Kehrein & S. Rabanus (eds.), *Language and Space. An International Handbook of Linguistic Variation. Volume 2: Language Mapping*, Berlin: de Gruyter, 433-457.

GOEBL, H. (2013a) "La dialectometrización del ALPI: rápida presentación de los resultados", in E. Casanova Herrero & C. Calvo Rigual (eds.), *Actas del XXVI Congreso Internacional de Lingüística y de Filología Románicas (Valencia 2010)*, vol. VI., Berlin / Boston: de Gruyter, 143-154.

GOEBL, H. (2013b) "La dialectometrització dels quatre primers volums de l'ALDC", *Estudis Romànics*, 35, 87-116.

GOEBL, H., C. SCAPOLI, S. SOBOTA, E. MAMOLINI, A. RODRIGUEZ-LARRALDE & I. BARRAI (2005) "Surnames and Dialects in France: Population Structure and cultural evolution", *Journal of Theoretical Biology*, 237, 75-86.

HAAG, C. (1898) *Die Mundarten des oberen Neckar- und Donautales*, Reutlingen: Buchdruckerei Egon Hutzler.

ILG/RAG (2003) *Normas ortográficas e morfolóxicas do idioma galego*, Santiago de Compostela: ILG/RAG.

NERBONNE, J. (2006) Identifying linguistic structure in aggregate comparison, *Literary and Linguistic Computing*, 21, 463-76.

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/>

NERBONNE, J., W. HEERINGA & P. KLEIWEG (1999) "Edit Distance and Dialect Proximity", in D. Sankoff & J. Kruskal (eds), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison,* Stanford: CSLI Press, 5-15.

PICKL, S. & J. RUMPF (2012) "Dialectrometric Concepts of Space: Towards a Variant-Based Dialectometry", in S. Hansen, C. Schwarz, P. Stoeckle & T. Streck (eds.), *Dialectological and folk dialectological concepts of space*, Berlin: Walter de Gruyter, 199-214.

PICKL, S. (2013) "Lexical meaning and spatial distribution. Evidence from geostatistical dialectometry", *Literary and Linguistic Computing*, 28, 63-81.

PICKL, S., A. SPETTL, S. PRÖLL, S. ELSPAß, W. KÖNIG & V. SCHMIDT (2014) "Linguistic distances in dialectrometric intensity estimation", *Journal of Linguistic Geography*, 2, 25-40.

SARAMAGO, J. (2002) "Diferenciação lexical interpontual nos territorios galego e português (estudo dialectométrico aplicado a materiais portugueses do *ALPI* e a materiais galegos do *ALGa*)", in R. Álvarez, F. Dubert & X. Sousa (eds.), *Dialectoloxía e léxico*, Santiago de Compostela: Consello da Cultura Galega – Instituto da Lingua Galega, 41-68.

SCHRAMBKE, R. (2010) "Language and space: Traditional dialect geography", in P. Auer & J. E. Schmidt (eds.), *Language and Space: An International Handbook of Linguistic Variation*. *Volume I: Theory and Methods,* Berlin, New York: de Gruyter, 87-106.

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

SIBLER, P., R. WEIBEL, E. GLASER & G. BART (2012) "Cartographic Visualization in Support of Dialectology", *Proceedings of AutoCarto - International Symposium on Automated Cartography*, Columbus, Ohio, USA, 16 September 2012 - 18 September 2012. <http://www.cartogis.org/docs/proceedings/2012/Sibler_etal_AutoCarto2012.pdf>

SNOEK, C. (2014) "Review of Gabmap: Doing Dialect Analysis on the Web", *Language documentation & conservation* 8, 192-208.

SOUSA, X. (2006) "Análise dialectométrica das variedades xeolingüísticas galegas", in M.C. Rolão Bernardo & H. Mateus Montenegro (eds.), *I Encontro de Estudos Dialectológicos. Actas*, Ponta Delgada: Instituto Cultural de Ponta Delgada, 345-362.

SOUSA, X. (forthcoming) "Between standard and dialects: linguistic distance in modern Galician".

SZMRECSANYI, Benedikt (2014) "Methods and objectives in contemporary dialectology", in I.A. Seržant & B. Wiemer (eds.), *Contemporary approaches to dialectology: The area of North, Northwest Russian and Belarusian vernaculars (Slavica Bergensia 12),* Bergen: Department of Foreign Languages - University of Bergen, 81-92.

TRASK, L.R. (2000) *The dictionary of historical and comparative linguistics*, Edinburgh: Edinburgh University Press.

WIELING, M. & J. NERBONNE (2015) "Advances in Dialectometry", *Annual Review of Linguistics*, 1, 243-264.

Dialectologia. Special issue, VI *(2016), 191-221.*
ISSN: 2013-2247

FIGURES

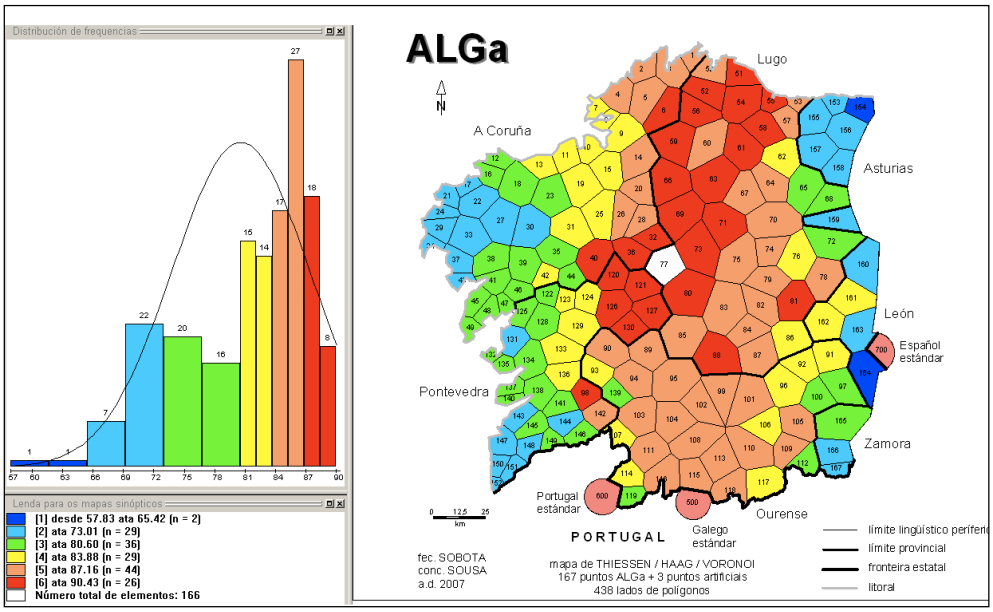

Figure 1. Similarity map of *ALGa* point L.28 (77), phonetic data (Dubert 2011).

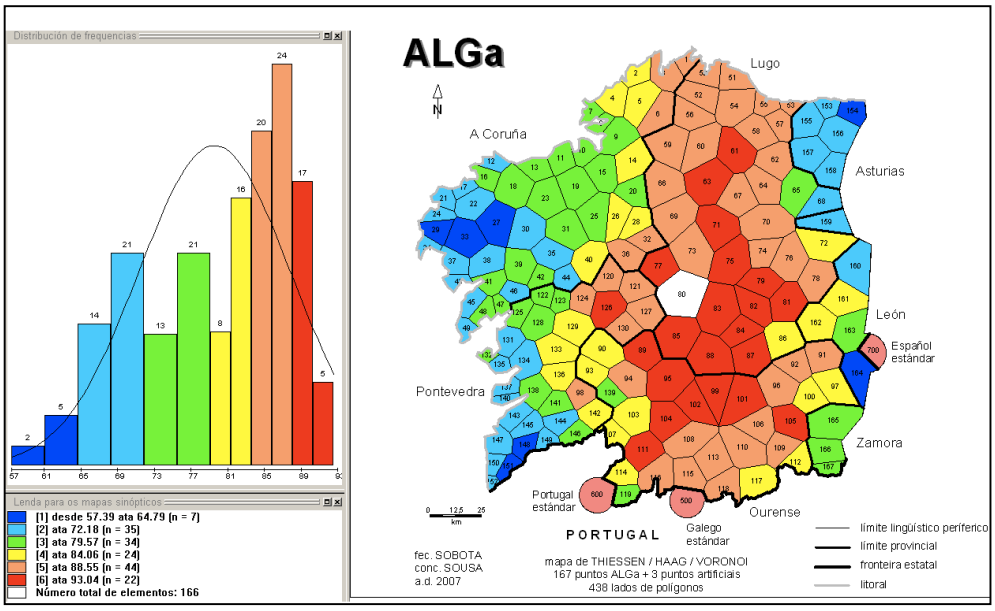

Figure 2. Similarity map of *ALGa* point L.31 (80), phonetic data (Dubert 2011).

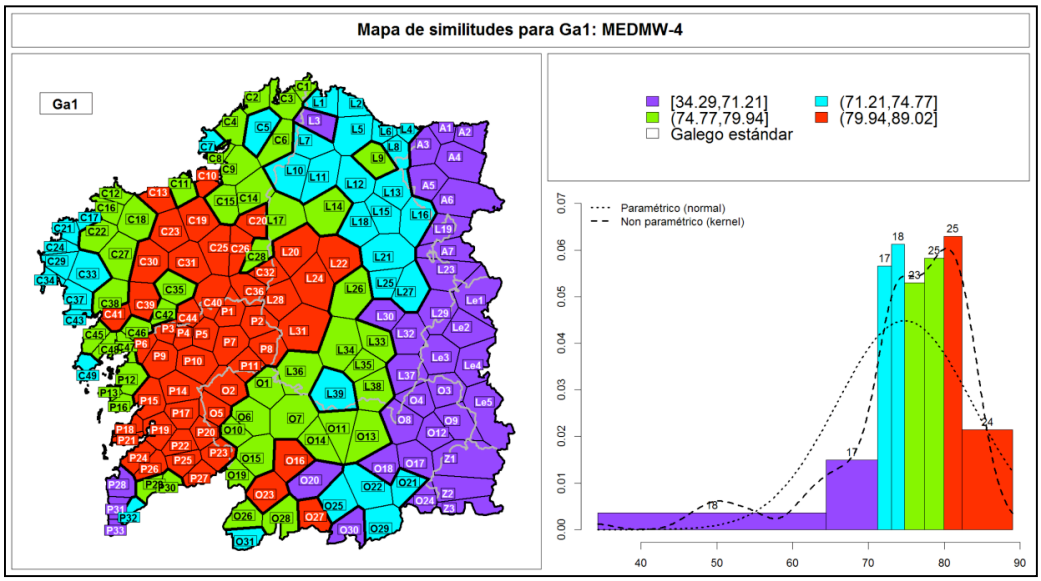

Figure 3. Similarity map of the standard variety, morphological data taken from the *ALGa* (Sousa forthcoming).

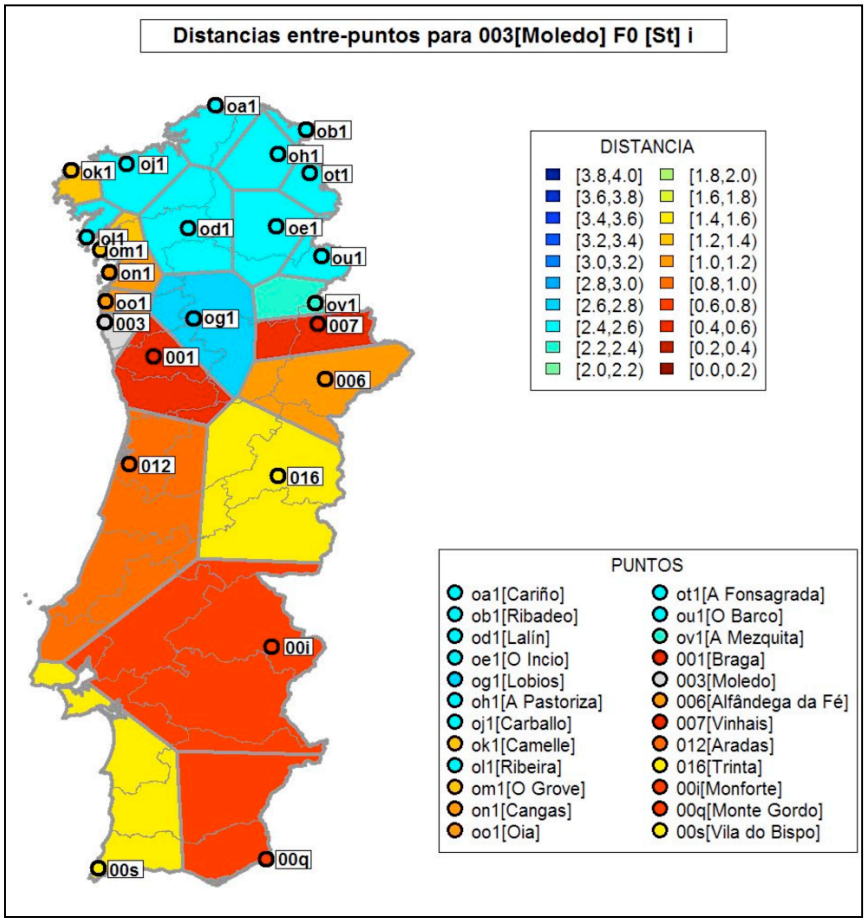

Figure 4. Similarity map of Portuguese point 003 (Moledo), data of intonation in polar question utterances (Fernández Rei, Moutinho & Coimbra 2014).

Dialectologia. Special issue, VI *(2016), 191-221.*
ISSN: 2013-2247

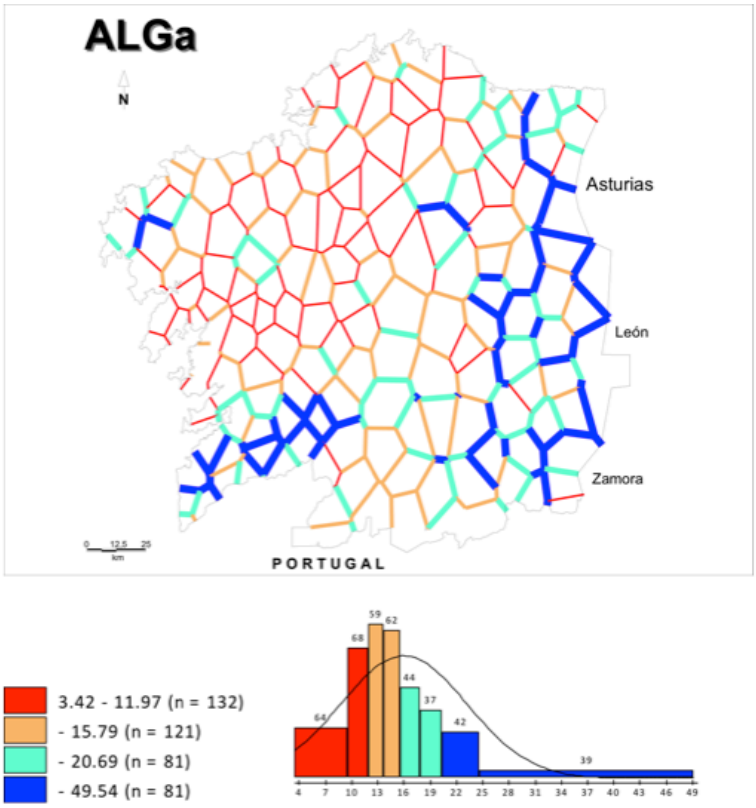

Figure 5. Honeycomb map of interpunctual differences, morphological data taken from the *ALGa* (Álvarez, Dubert-García & Sousa 2006).

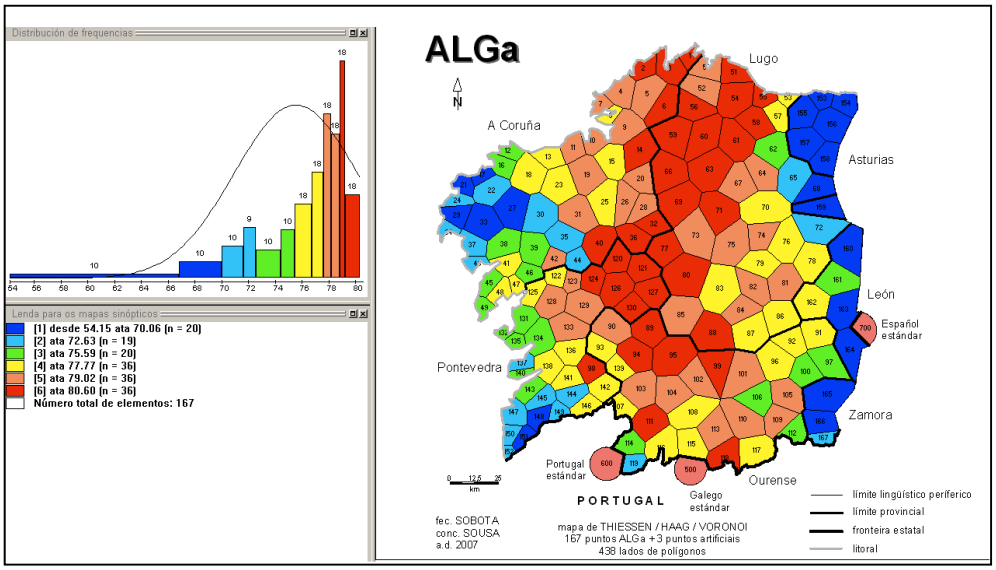

Figure 6. Choropleth map of the synopsis of the mean values of 167 similarity distributions, phonetic data taken from the *ALGa* (Dubert 2011).

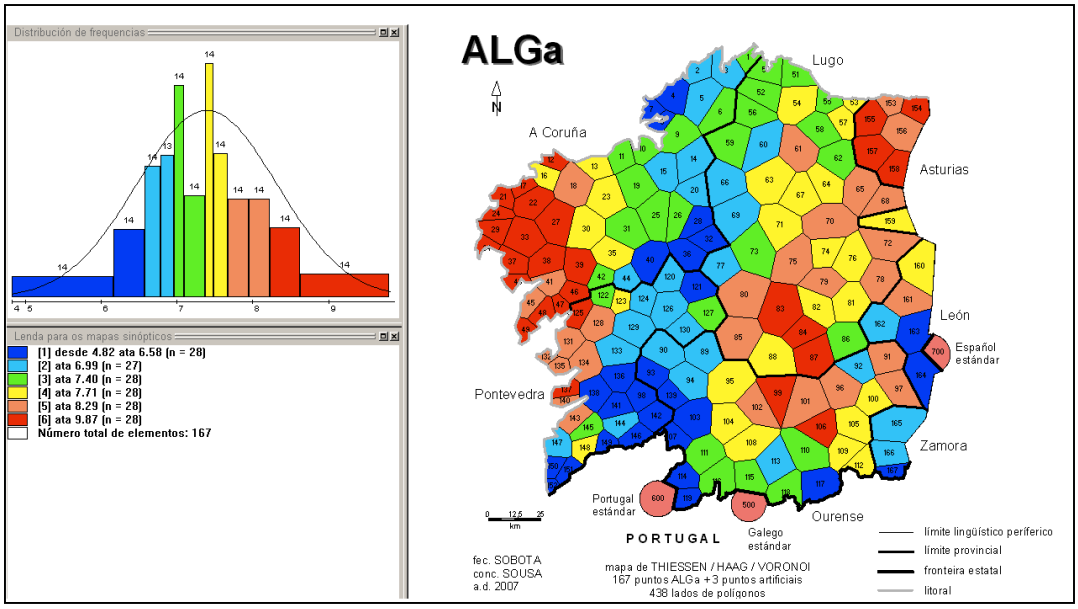

Figure 7. Choropleth map of the synopsis of the standard deviation of 167 similarity distributions, phonetic data taken from the *ALGa* (Dubert 2011).

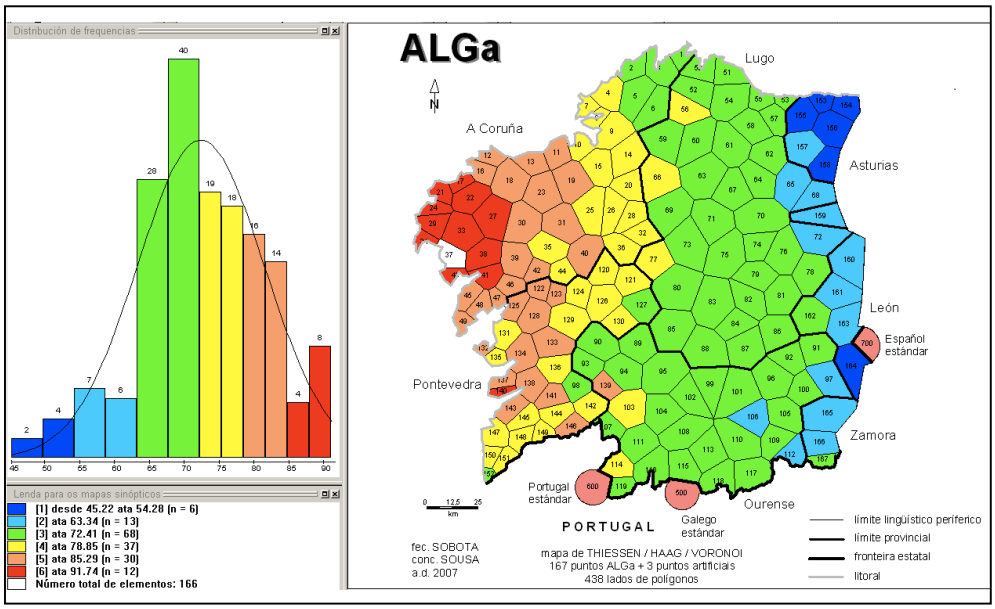

Figure 8. Similarity map of *ALGa* point C.37, phonetic data (Dubert 2011).

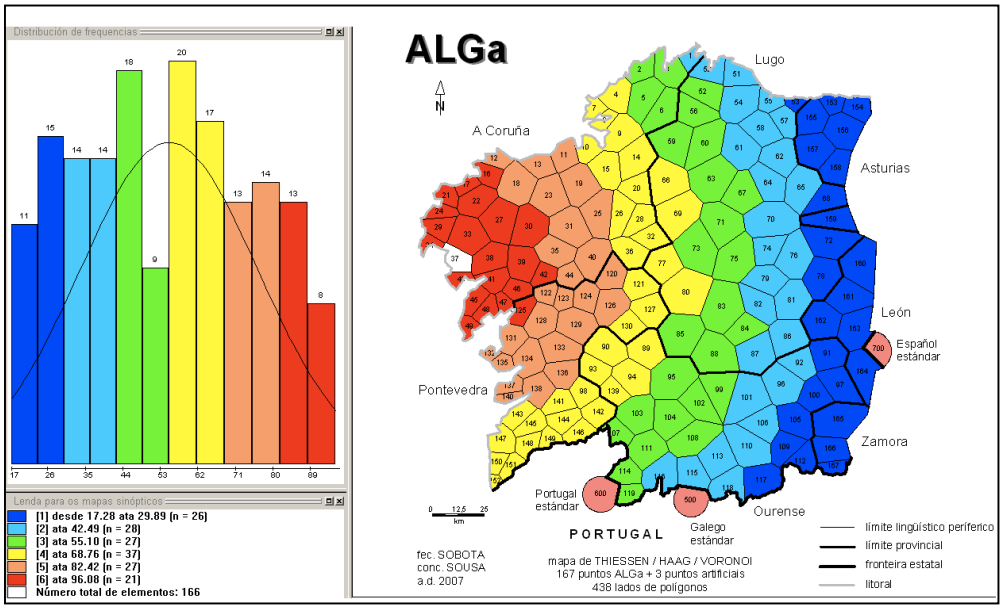

Figure 9. Proximity map of *ALGa* point C.37 (Dubert 2012, 2013).

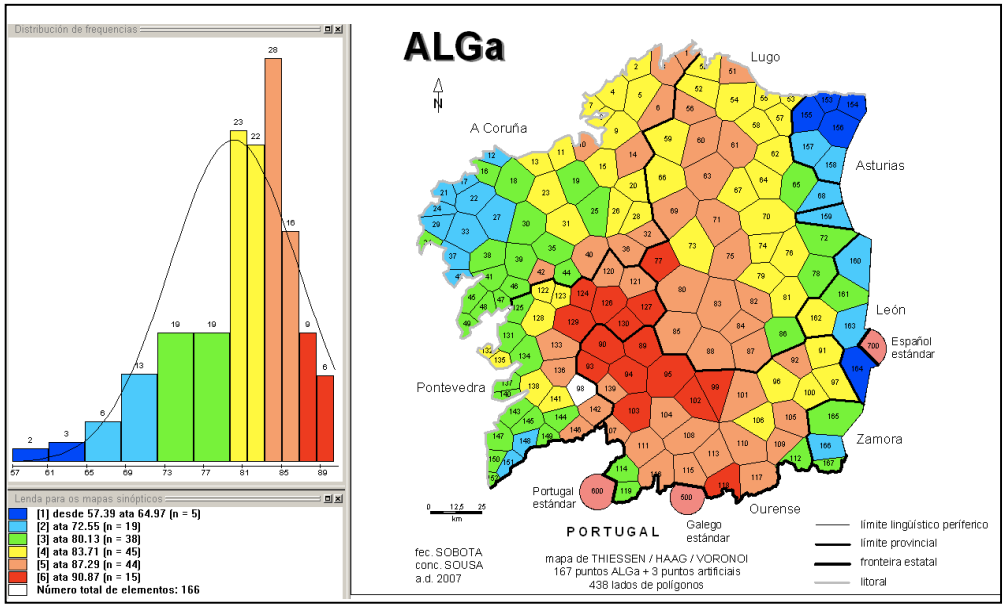

Figure 10. Similarity map of *ALGa* point O.10 (98), phonetic data (Dubert 2011).

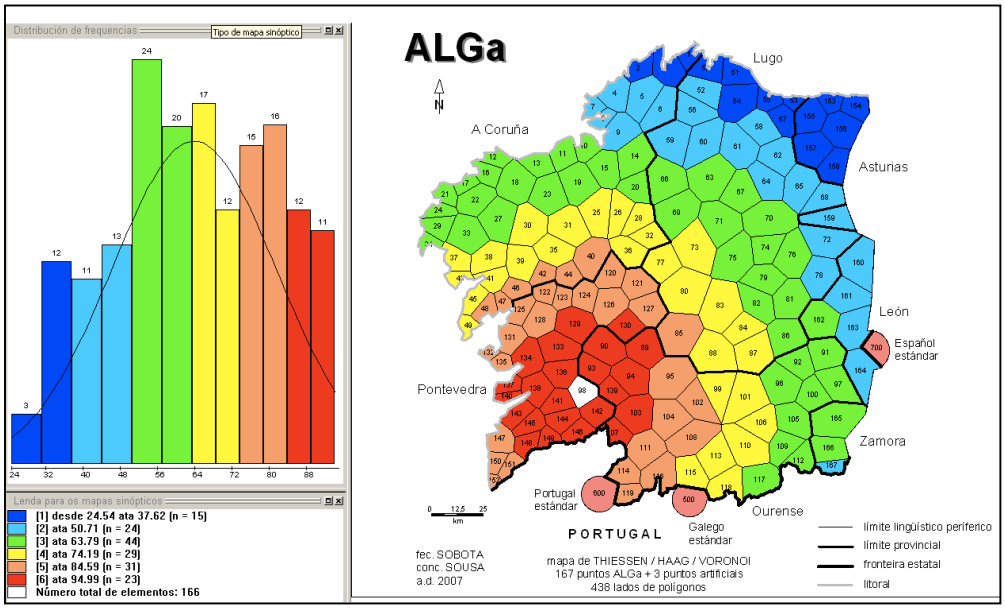

Figure 11. Proximity map of *ALGa* point O.10 (98) (Dubert 2012, 2013).

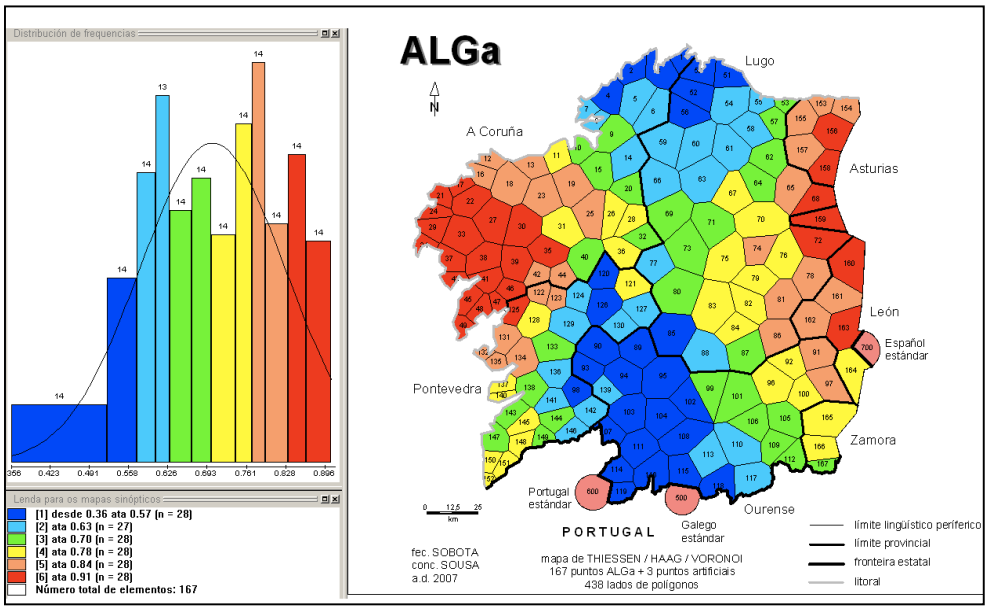

Figure 12. Choropleth map of the correlation values between 167 similarity values and 167 proximity values, phonetic data taken from the ALGa (Dubert 2012).

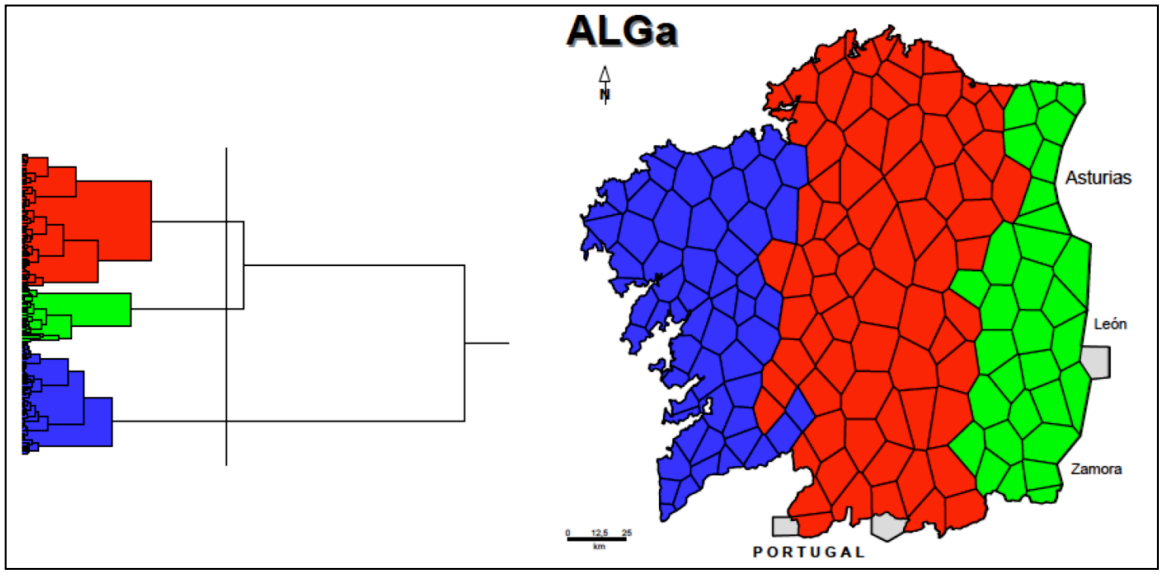

Figure 13. Dendrographic classification of 167 dialectological objects (ALGa-points) and spatial conversion of the tree (Sousa 2006).