Geographical proximity and linguistic similarity. A correlative analysis of Galician phonetic data[1]

Francisco Dubert-García
Universidade de Santiago de Compostela
francisco.dubert@usc.es

Keywords: Correlative Dialectometry, Galician, ALGa

1. Introduction

In 2006 Rosario Álvarez, Francisco Dubert and Xulio Sousa published the first comprehensive dialectometric study of Galician speech varieties on the basis of data from the *Atlas Lingüístico Galego* (*ALGa*) (Álvarez, Dubert & Sousa 2006). For that study, the authors drew on a database of 321 working maps: 100 contained phonetic, 121 morphological and 100 lexical data. The study demonstrated that Galician dialects fall into two major areas: a western coastal strip and the remainder of the language's territory, respectively.

Some years later, Dubert (2011) published a new work including 230 phonetic working maps. That study presented the difficulties for designing a database and carried out an analysis with the classical tools of Salzburg Dialectometry (similarity profiles, synopses of the values of skewness and standard deviation, a honeycomb map and a dendrographic classification). I would now like to present an application of Correlative Dialectometry (Goebl 2005, 2006, 2008) to the same phonetic data. This tool makes it possible to check the degree of correlation between two distributions of data. In this case, I will analyse the correlation between the similarity indices of the 230 phonetic working maps referred to above and the geographical distance between points.

To achieve this goal, in §2 I will present some basic assumptions, in §3 the correlative analysis and in §4 some conclusions.

2. Some caveats on the notion of linguistic area

Linguistic areas are not the same thing as dialects or languages: linguistic areas are geographical entities, somehow represented on maps; dialects, on the contrary, are linguistic systems. *Linguistic areas* are *geographical spaces* delimited by means of linguistic features. Most commonly linguistic areas are established from a qualitative point of view. *Qualitative linguistic areas* are a kind of *geographical areas* determined by the territorial extension of one or more linguistic features. Thus in Galicia we observe a linguistic area constituted by the geographical extension of the *gheada*: the places where an old /g/ evolved to /x, h/. The phenomenon known as *seseo* gives rise to another different linguistic area: the territory where a dental fricative /θ/ was not developed from a laminal /s/. Three

[1] This work has been carried out with funding from the Ministerio de Ciencia e Innovación of the Spanish Government, under the project "Estudio perceptivo de la variación prosódica dialectal del gallego" (FFI2009-12738). I wish to thank Xulio Sousa, Xosé Afonso Álvarez, and João Saramago for their comments on previous drafts.

Álvarez Pérez, Xosé Afonso, Ernestina Carrilho & Catarina Magro (eds.) (2012) *Proceedings of the International Symposium on Limits and Areas in Dialectology (LimiAr). Lisbon, 2011*. http://limiar.clul.ul.pt. Lisboa: Centro de Linguística da Universidade de Lisboa. 71-91.

ISBN 978-972-96404-9-0

different areas result from conflating the geographic extension of both features: a) an area with *seseo* and *gheada*; b) an area with *gheada* but not *seseo*; c) an area without either *gheada* or *seseo*.

But linguistic areas can also be established from a quantitative point of view. *Quantitative linguistic areas* are *geographical areas* obtained by, for example, grouping together places whose dialects share high indices of similarity. I use *indices of similarity* in the sense of Salzburg Dialectometry (Goebl 2008): an index obtained by counting the similarities and differences displayed by one point of the network vis-à-vis the others. Red areas in Figures 3, 4, 5, 6 and 8 represent quantitative linguistic areas obtained by grouping together points with high indices of similarity.[2]

The dialects spoken within a linguistic area are not uniform. From a qualitative viewpoint, some localities are included in a linguistic area because their dialects share specific features even though they may differ in others: for example, the area defined by *gheada* contains dialects both with and without *seseo*. From a quantitative point of view, some localities are included in a linguistic area because they share a range of values; however, similar values may be reached by sharing different features.

Quantitative linguistic areas are *relative*, not *absolute*. They are the output of several factors, such as (1) the way taxatation is carried out (i.e. the criteria used to create the working maps); (2) the number and kind of working maps included in the database; (3) the number of points that conform the network. For example, the database that holds the choropleths I will present was elaborated from 230 working maps of phonetic data (Dubert 2011). 230 working maps are a *sample* of the language, not to be confused with the language *per se*.

The maps were taken from the *Atlas Lingüístico Galego* (*ALGa*), a survey carried out between 1974-1976 in 167 localities: 152 in Galicia, 7 in Asturias, 5 in León and 3 in Zamora. If, for example, instead of 15 localities located outside Galicia, the designers of the survey had included 20 such points, the overall indices of all the points would be affected.

On the other hand, most possible working maps with a clear dialectal profile, such as that of "Magosto" 'chestnut roasting',[3] in Figure 2, were excluded from the database because they lack a significant number of answers; their inclusion might have introduced noise in the statistical calculations, since the VDM program[4] that I use to compile the choropleths and carry out the mathematic calculations counts blanks as a different answer. Other potential working maps were excluded because of the high number of Castilian loanwords: this is the case of "Martelo" 'hammer', where the Spanish loanword *martillo* was obtained in most points, contaminating the original geographical distribution of [e] and [ɛ] as possible Galician outcomes of Latin ĕ. If we had the answers and complete working maps for both these maps (and others), we might have introduced them in the database and, as a consequence, the resulting indices would be different again. In fact, the statistical indices are only reflexes of

---

[2] Indices of similarity like those of Figures 3, 4, 5 etc. are not the only source of quantitative areas. Dialectometric methods (Goebl 2006) allow us to detect different kinds of quantitative areas obtained by mapping the values of statistical indices such as arithmetic means of similarity, standard deviation of the means, and skewness of the distribution; these are known as parameter maps and each map has its own linguistic interpretation. Correlative maps are one kind of parameter map.

[3] "Magosto" (*ALGa III*, map 12), is a fire used to grill chestnuts and also refers to the party where the chestnuts are eaten. Black triangles represent *mag*[u]*sto*, black circles represent *mag*[o]*sto*, white circles represent non-cognate answers. Points without a symbol have no answer.

[4] The choropleths presented in this work were created with a program called VDM (*Visual DialectoMetry*), developed by Edgar Haimerl around 1997-2000. The focus of VDM is numerical (the rapid execution of many statistical computations) and graphic (the optimal visualisation of the numerical results, Goebl 2006: 413).

the material put into the database.

Quantitative dialectology is also relative in its outcomes. For instance, the indices of similarity of each point are relative to the point in question: in the Figure 3, points 77 (L.28 in the *ALGa* network) and 80 (L.31 in *ALGa*) are grouped together in the same red area; however, as the line drawn in Figure 4 shows, 77 and 80 display different patterns across the territory, because each point has its own profile of similarity and the localities linguistically closer to it (the red ones) are different.

The algorithms used to produce the visualization of the choropleths (how many cuts are introduced into the continuum of row indices and, as a consequence, how many groupings are obtained) or those used to calculate the indices of similarity also determine the final form of the choropleths: comparison between Figures 5 (with 4 groupings) and 6 (the same data with 6 groupings) illustrates this question.

So, linguistic areas established on quantitative grounds strongly depend on the method of elaborating and managing the database.

Finally, quantitative linguistic areas do not need to agree with borders between historical languages or dialects. The proposal of a quantitative analysis is grouping points that share some range of indices; the linguistic variety *V.1*, customarily assigned to the language *L.A*, may share higher indices of similarity with the variety *V.2*, customarily assigned to the language *L.B*, than with the variety *V.3*, customarily assigned to language *L.A*. For example, in my database:

- Le.4 (*V.1,* Galician *L.A*) and Le.5 (*V.2,* Leonese *L.B*) share 71.3% similarity, however
- Le.4 (*V.1, L.A*) and C.29 (*V.3, L.A*) share 51.74% similarity.

So, in this paper you will look at quantitative, relative, geolinguistic areas which are the outcome of a process of comparison of features obtained in a network. These areas, obtained from a database that is a sample, should not be confused with dialects, languages or any kind of systematic linguistic varieties. The geolinguistic areas will only show the way speakers handle certain linguistic features or bundles of features, or in the words of Goebl (2006: 412):

> The Salzburg research team defines the main aim of DM [Dialectometry] in the exploration of the 'basilectal management of space by the HOMO LOQUENS'. This basic theoretical position assumes that geographical 'space' is a texture of complex relations in which man evolves and settles by means of speech. Thus, human speech is considered as being *variable* and space as *invariable*.

The areas are traces of the linguistic interactions and exchanges carried out by speakers across the territory. Quantitative areas are immanent structure hidden in the row data. And these areas are waiting to be discovered and explained.[5]

[5] "Le strutturazioni variegate delle carte di un qualsiasi LA [Linguistic Atlas] rappresentino delle emanazioni speciali di un sistema stocastico sottostante retto da un gran numero di interdenpendenze, correlazioni e connessioni" (Goebl 2008: 26-27). Dialectometrical analysis leads to the study and discovery of regularities and laws internal to the system.

3. On the correlation between linguistic similarity and geographic proximity

The geographical spread of a linguistic feature is determined by facts external to the language; in fact, the distribution of individual linguistic features or bundles of features constitutes good evidence for the existence of historical (sociological, economic etc.) connections between the localities included in the area covered by those features: these are areas where people interact easily (Viejo (2003) explores this possibility for the Asturian language). Individual features and bundles of features travel and spread out across the territory and may be exchanged through conversation. In fact, features are used to produce what Goebl (2005) calls *symbolic demarcation* of speakers and Le Page & Tabouret-Keller (1985) call *acts of identity*: those acts of identity are the cause of sharing and spreading features.

We can think of linguistic areas sharing high indices of similarity as clusters of localities that maintain common social interactions and contacts: similarities are evidence for interaction and the interactions explain the similarities. It is through speaker interaction that linguistic features spread out across a territory, whereby linguistic areas and similarities emerge as a consequence.

One factor that promotes or blocks social interaction and the realization of acts of identity is geographical distance, because speakers need to keep themselves in touch in order to perform acts of identity. However, although geographical contiguity is a requisite for the maintenance of contact, geographical contiguity does not necessarily entail social interaction or linguistic exchange: Figure 1 shows that geographically contiguous localities may have low indices of similarity, sharing fewer features than expected. Figure 1 is a dendrogram with 7 clusters and its correspondent chorem (the map), both obtained from the aforementioned 230 phonetic working maps (Dubert 2011). These graphic representations arise from grouping the points of the network by internal similarities.[6] At first glance, this seems to resemble a classical distribution of linguistic areas spreading across the territory; however, observing the dendrogram, we see that the plain distribution of the coloured areas has a deep, hierarchic, structure. Area 5 is a sister of 6; and the cluster [5, 6] is a sister of 7. On the other side, area 2 is a sister of 1; the cluster [1, 2] is a sister of 3; and [[1, 2], 3], a sister of 4; finally, the group [[[1, 2], 3], 4] is a sister of [[5, 6], 7]. So, although areas 5 and 2 are geographically contiguous, with points in geographical contact, they are only indirectly related in geolinguistic terms: 5 and 2 pertain to different branches (or dialectal types, or linguistic macro-areas) and they are *immediate constituents* of different clusters.

If, then, geographical contiguity does not imply linguistic similarity, other factors must be involved in the spread and creation of linguistic areas. However, in other points we find a good correlation between linguistic similarity and geographical proximity. If we compare the choropleths contained in Figures 6 and 8 we see two different similarity profiles.

Figure 6 is a choropleth of the similarity distribution of C.37, a north-western point of the *ALGa* network. Note how an increase in geographical distance is accompanied by a decrease in linguistic similarity (warm colours mean high indices of similarity, cold colours low indices of similarity). However, in Figure 8 we see a choropleth of the similarity profile of O.10, a

[6] "By a successive fusion or agglomeration of (respectively) two most similar elements of the similarity matrix [...] a binary hierarchy of classes or clusters (or 'dendremes') is generated. In this process, the (bigger) dendremes located near the root of the tree have a greater inner (numerical) heterogeneity than the (smaller) dendremes near the leaves" (Goebl 2006: 420-21).

southern point in the network. In this case, the similarity profile obtained naturally decreases in a smooth slope towards the East and North, but plunges abruptly to the West.

In Figure 7 we see the profile of geographical proximity of C.37 and in Figure 9 the same profile of O.10. The colours are selected according the degree of proximity, ranging from red for maximal proximity to dark blue for minimal proximity. Actually, both choropleths represent the geometrical (Euclidian) distance *between points on the map*, not geographical distances between the localities; however, this Euclidean distance represents, on a scale, the geographical distances existing between the localities of the network.[7] It is noticeable that the profiles of linguistic similarity and geographic proximity obtained in C.37 (Figures 6 and 7, respectively) do not match exactly, but they do match better than the profiles of O.10 (Figures 8 and 9).

As Figures 10 and 11 show, we can use scatter graphs to display the values that every point reaches in two distributions, Tables 1 and 2. The first column of each Table contains the indices of geographical (in fact, Euclidean) proximity between one point P and the others; the second column contains the indices of linguistic similarity between P and the others. Both Tables are ordered by proximity. In Table 1, P corresponds to C.37 and in Table 2 P corresponds to O.10. The scatter graphs of Figures 10 and 11 enable us to check at first glance the degree of correlation that the two distributions reach in C.37 and O.10, respectively (which is difficult to see in Tables 1 and 2). Indices of geographical distance, taken as the independent variable, are plotted along the horizontal axis; indices of linguistic similarity, considered the dependent variable, are plotted along the vertical axis. Figure 10 shows that C.37 achieves a good correlation: similarity increases as proximity increases; whereas Figure 11 shows that O.10 does not show so good a correlation: there are a good number of points where similarity seems to be autonomous of proximity. In fact, Figure 11 shows a lot of points (those located in the left half of the highest row) with indices of similarity approximately ranging between 80% and 90%, despite the fact that these points are relatively distant from O.10, as shown by their low indices of proximity (ranging approximately from 30 to 50):

- O.10 and L.1: proximity index 33.21, similarity index 83.48
- O.10 and C.1: proximity index 32.80, similarity index 84.78
- O.10 and L.4: proximity index 31.57, similarity index 81.66

- O.10 and P.22: proximity index 94.99, similarity index 80,79
- O.10 and P.23: proximity index 94,92, similarity index 83.89
- O.10 and P.25: proximity index 93.09, similarity index 75.55

4. Correlative dialectometry

Goebl (2005: 327-332) proposes to reduce each pair of distributions like those of Tables 1 and 2 to a coefficient that permits comparison between the correlations achieved by each point. The statistical tool that produces the coefficient is known as the *Bravais-Pearson correlation coefficient*. The coefficient's values range from -1 to +1. If the value obtained

[7] I wish to thank Professor Goebl for providing me with the file GeoProx.vdd, which creates the profiles of Euclidean (geometric) distance for each point of the *ALGa*'s network and which also generates the indices for measuring the distance. These indices are those really used to measure the distance.

is -1, we have a perfect inverse linear relation: when Y (proximity) increases, X (similarity) decreases proportionally; if the value is 0 we have no linear correlation between X and Y, and proximity and similarity seem to be autonomous; when the value is +1, we find a direct linear relation: when Y increases (proximity), X (similarity) increases proportionally. C.37 has a value of 0.91 (the highest value in the sample), which means that in C.37 there exists a good correlation between linguistic similarity and geographic proximity; however, in O.10 the value obtained was 0.36 (the lowest in the sample), which means that some degree of autonomy exists between proximity and similarity.

We could calculate the value of the correlation coefficient for each point and plot the corresponding scatter graph. However, as geolinguists, we are compelled to see this information on a map, because, as Goebl (2005: 323) humorously says "extra mappas nulla salus dialectometrica sive geolinguistica" (without maps, no salvation in dialectometry or geolinguistics):

> Les phénomènes dont la géographie linguistique s'occupe depuis plus de cent ans, ne peuvent se manifester et dérouler que dans l'espace. Or, vu les conditions matérielles et les contraintes psychologiques que sont les nôtres, il est de toute première importance que les géolinguistes reportent les faits observés par eux sur des schémas iconiques et qu'ils s'y réfèrent, en permanence, dans leurs réflexions et discussions. Qui, en tant que géolinguiste, renonce à l'usage de cartes ou finit même par les répudier, anéantit de ce fait sa propre discipline (Goebl 2005: 323).

As we have 167 pairs of distributions of proximity and similarity, we have 167 values for the Bravais-Pearson correlation coefficient. VDM allows us to calculate each coefficient and represent the outcomes in a choropleth.

Figure 12 shows the geographical distribution of all the values obtained with the Bravais-Person coefficient: the array of 167 coefficients is segmented and grouped in six different intervals according to their values: three intervals for coefficients below the mean and three for those above the mean. Warm colours are assigned to values above the arithmetic mean (red is highest) and cold colours are used for values below the mean (dark blue is lowest).

Looking at the histogram, the first thing we notice is that:

a) The values of coefficients range from 0.36 to 0.91;
b) No zero or negative coefficients are found; only positive ones;
c) 159 points out of 167 (94.08%) have coefficients ranging between 0.50 and 0.91;
d) 89 points out of 167 (53.29%) have coefficients ranging between 0.70 and 0.91.

Thus we observe in general a strong tendency toward direct correlations, with higher correlation coefficients than in other linguistic domains. As shown by Goebl (2005, 2006 and 2008), the values for Gallo-Romance range between -0.25 and 0.91; for English, between -0.43 and 0.89; for Italo-Romance, between -0.44 and 0.86.[8]

Another salient feature of the choropleth is a well-established distribution pattern of data following vertical bands, from north to south: warm colours appear in the West, followed by cold colours in the centre, then warm colours again in the East. This west-to-

[8] The correlations for Gallo-Romance are calculated from 1117 phonetic working maps; those for Italo-Romance from 1630 phonetic working maps; those for English from 597 lexical and morphological working maps.

east patterning is similar to other qualitative and quantitative patterns found in Galician dialects.

The pattern of Figure 12 is so well ordered and delimited that it cannot have arisen by chance: we find compact clusters of red and dark blue areas and transitions between the clusters, not a *totum revolutum* with unordered value sets, warm and cold colours intermingled or appearing in disarray. The meaning of the colours is explained by Goebl (2008: 99-101):

> Les zones rouges correspondent à des régions où l'étalement des similarités linguistiques dans l'espace obéit dans une grande mesure aux impératifs euclidiens de ce dernier. Ceci signifie que, sur une carte de similarité quelconque, le taux de la similarité linguistique décroît, avec l'éloignement du point de référence respectif, d'une façon plus ou moins régulière (c'est-à-dire statistiquement linéaire). L'inverse est vrai pour les zones teintées de bleu. C'est ici que l'étalement des similarités linguistiques dans l'espace a été libéré complètement des contraintes euclidiennes de l'espace, évidemment pour des raisons sociales et politiques de toute sorte. Ceci signifie que l'échelonnement spatial des similarités linguistiques peut ou bien dépasser les impératifs de la géométrie euclidienne ou bien rester en retrait par rapport à ces derniers. Ces écarts peuvent survenir à la suite d'un rayonnement culturel et social dans le premier des deux cas, ou représenter la conséquence d'un isolement du même ordre dans l'autre.

So, in the red areas we find indices of correlation between language similarity and spatial proximity closer to 1, while in the dark blue areas we find indices further removed from 1. Therefore the red areas are close to direct correlation: similarity proportionally increases with proximity or, to put it the other way around, linguistic difference increases with geographic distance. In the areas with cold colours there is greater autonomy between proximity and similarity. Looking again at Figures 7 and 8, we find verification of what Goebl predicted: C.37 (Figure 6) has a pattern in which "le taux de la similarité linguistique décroît, avec l'éloignement du point de référence respectif, d'une façon plus ou moins régulière (c'est-à-dire statistiquement linéaire)"; we cannot say the same for O.10 (Figure 8), with its abrupt fall in the west and its northward lengthening.

But we should recall the caveats given at the beginning of this paper: the correlation shown in Figure 12 depends on the material introduced in the database.

Interestingly, this distribution is coherent with previous analyses (Dubert 2011). Figure 13 shows the distribution of arithmetic means of similarity. To create this choropleth, we calculate the mean index of similarity for every point. The choropleth shows that high means appear in the central area (warm colours in Figure 13), implying that these points achieved high indices of similarity; here we also find low indices of correlation (cold coloured areas in Figure 12): similarity is obtained to such a degree that it ignores geographic distance. The areas with high indices of correlation (red colours in Figure 12) partially overlap areas with low indices of means (cold coloured areas in Figure 13; in fact, the areas with low means are larger and occupy all the east and all the west); so the red areas in Figure 12 (with high indices of correlation) are *located at extremes of the domain and have low arithmetic means*.

If we now look at Figure 14, we shall see the distribution of values of standard deviation. Warm colours group together points with high indices of standard deviation; cold colours,

points with low indices. This choropleth must be interpreted in the light of that in Figure 13: points with high indices of means have low standard deviation. However, points with low means (cold colours in Figure 13) can be obtained with greater disparity within the sample (Dubert 2011). Lower means may arise in two ways:

a) Greater dispersion of values around the arithmetic mean: some points share many features with others and few with the rest; this distribution has indices far from the mean, because of their similarity or difference; so there are points that are very similar to each other and very different from the rest: in this case, we find high indices of standard deviation.
b) Little dispersion around the arithmetic mean: a point can have low indices of similarity with all the rest of the points, with many indices of similarity around the low mean; so the arithmetic mean reflects the fact that the point is uniformly different from or similar to the rest; in this case we get low indices of standard deviation.

Now, if we compare Figures 12, 13 and 14, we see that the points *located at the extremes of the network* with high correlation indices (red colours in Figure 12) have roughly low means (dark blue colours in Figure 13), and high indices of standard deviation (red colours in 14). Thus dialects in the same area with low means and high standard deviation have well-defined personalities: the north-western dialects share a lot of features internally but few with dialects outside their group; and the same may be said of the north-eastern dialects.

The dialects of central Galicia tend to show high means of similarity and low standard deviation: they are uniformly similar, because they share a lot of features, regardless of distance. The histogram in Figure 6 shows 79 points (warm colours) with similarity indices above the mean (72.61) and 87 (cold colours) with similarity indices below the mean; the histogram of Figure 8 shows 104 points (warm colours) with similarity indices above the mean (80.13) and 62 (cold colours) below the mean. Thus in O.10 (Figure 8) we find more similarity.

In general, the dialects of south-eastern and south-western territories (also in the extremes of the network) tend to have low means and low indices of standard deviation: they are uniformly different among themselves and also from the rest, regardless of distance. Note that the south-western and south-eastern dialects are just as geographically peripheral as the north-western and south-eastern ones: they are all located on the edges of the network; however, only in the north-western and north-eastern strips do we find correlation between proximity and distance (similarity is highly dependent on proximity), while in the rest (including the other marginal areas), similarity is more independent of proximity.

4. Discussion

An explanation for the Galician linguistic data must refer to the economic, geographical, social and historical background of the Galician-speaking area: this is where Dialectology and History intersect.

To explain the distribution of correlation values, we need to recall that:

- Along the North of the Iberian Peninsula, most isoglosses run from north to south.
- The north-western dialects are isolated: this really is a *finis terrae*. These dialects are

only in contact with those spoken to their east. In the north-western dialects, high correlations between proximity and similarity seem to accord with isolation and conservatism and with their peripheral position in the network.

- The north-eastern dialects are not isolated: they are contiguous to the west with the rest of Galician dialects, and to the east with western Astur-Leonese dialects. However, the sharing of features with Astur-Leonese makes the Galician north-eastern dialects more different. They also occupy a peripheral position in the network.
- The central dialects are very intermingled, sharing a lot of features regardless of distance; this probably makes similarity more independent from proximity, because we find similarity across distance. Low correlations accord with the sharing of features across this area.
- Southern eastern and southern western dialects seem to show autonomy between proximity and similarity, although they occupy peripheral positions.

It seems, then, that areas with indices of correlation closer to 1 correlate roughly with geographically peripheral areas with low means of similarity and high indices of standard deviation.

I must stress once more that what these analyses identify are not dialects or boundaries between languages or dialects. Rather, they allow us to infer how speakers treat linguistic features. It matters little whether the row data are qualitative or quantitative: in qualitative geolinguistics, we are witnessing the exchange of specific features; in quantitative geolinguistics, the exchange of unordered sets of features. In either case, we see where linguistic interactions are intense enough to produce exchanges. The sharing features and bundles of features helps to establish a variety of areas, to account for which we must look to the historians of the Galician language.

References

ALGa III = Instituto da Lingua Galega (2000) *Atlas Lingüístico Galego. Volume III. Fonética.* A Coruña: Fundación Pedro Barrié de la Maza.

Álvarez, Rosario, Francisco Dubert García & Xulio Sousa Fernández (2006) "Aplicación da análise dialectométrica aos datos do *Atlas Lingüístico Galego*". In Rosario Álvarez, Francisco Dubert García & Xulio Sousa Fernández (eds.) *Lingua e territorio*. Santiago de Compostela: Instituto da Lingua Galega & Consello da Cultura Galega. 461-493.

Dubert García, Francisco (2011) "Developing a database for dialectometric studies: the ALGa phonetic data. Dialectometrical analysis of 230 working maps". *Dialectologia et Geolinguistica* 19. 23-61.

Goebl, Hans (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. 321-367.

Goebl, Hans (2006) "Recent advances in Salzburg Dialectometry". *Literary and Linguistic Computing* 21(4). 411-435.

Goebl, Hans (2008) "La dialettometrizzazione integrale dell'AIS. Presentazione dei primi risultati". *Revue de Linguistique Romane* 285-286. 25-113.

Goebl, Hans (2011) "Areas, fronteras, similitudes y distancias: lección breve de geolingüística cuantitativa". In Ramón de Andrés Díaz (ed.) *Lengua, ciencia y fronteras.* Uviéu: Trabe. 11-34.

Le Page, Robert Brock & Andrée Tabouret-Keller (1985) *Acts of Identity.* Cambridge: Cambridge University Press.
Viejo, Xulio (2003) *La formación histórica de la llingua asturiana.* Uviéu: Trabe.

Appendix

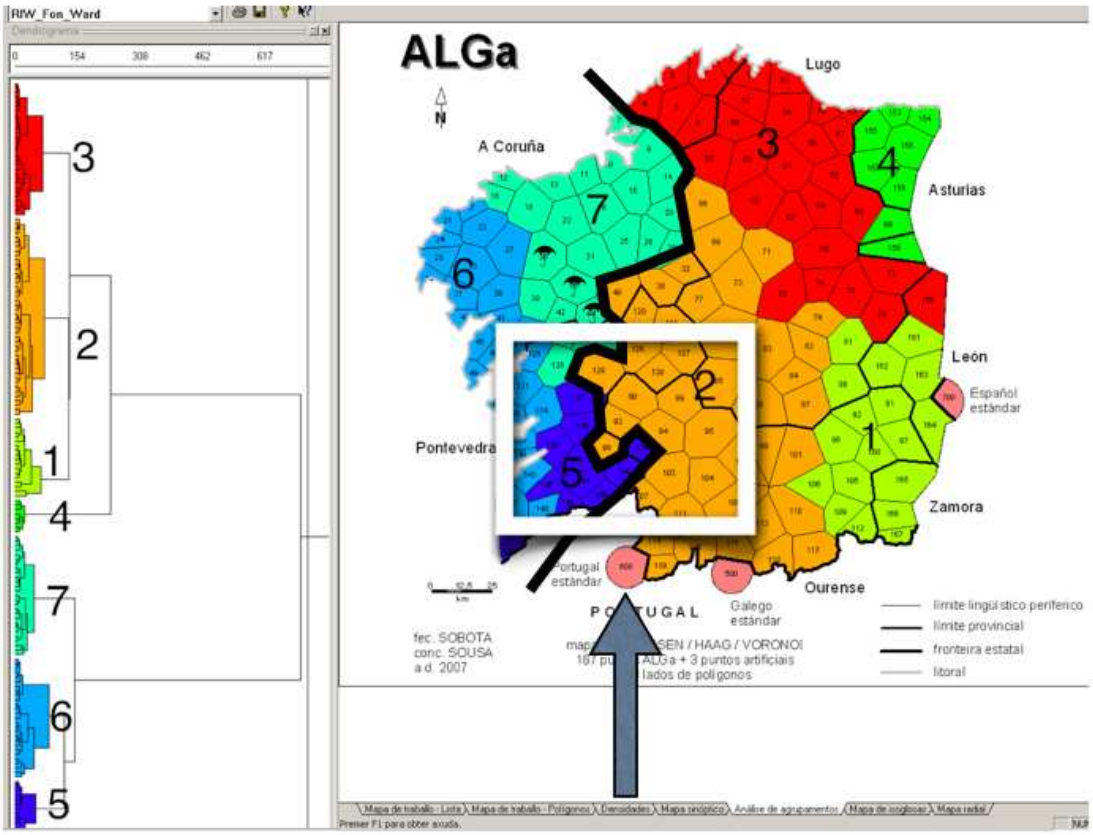

FIGURE 1. Dendrographic classification of the 167 points in *ALGa* with phonetic data and the corresponding chorem. The umbrella in 30, 35 and 44 means that the indices of these points are not very reliable because these points have a lot of blanks in the database

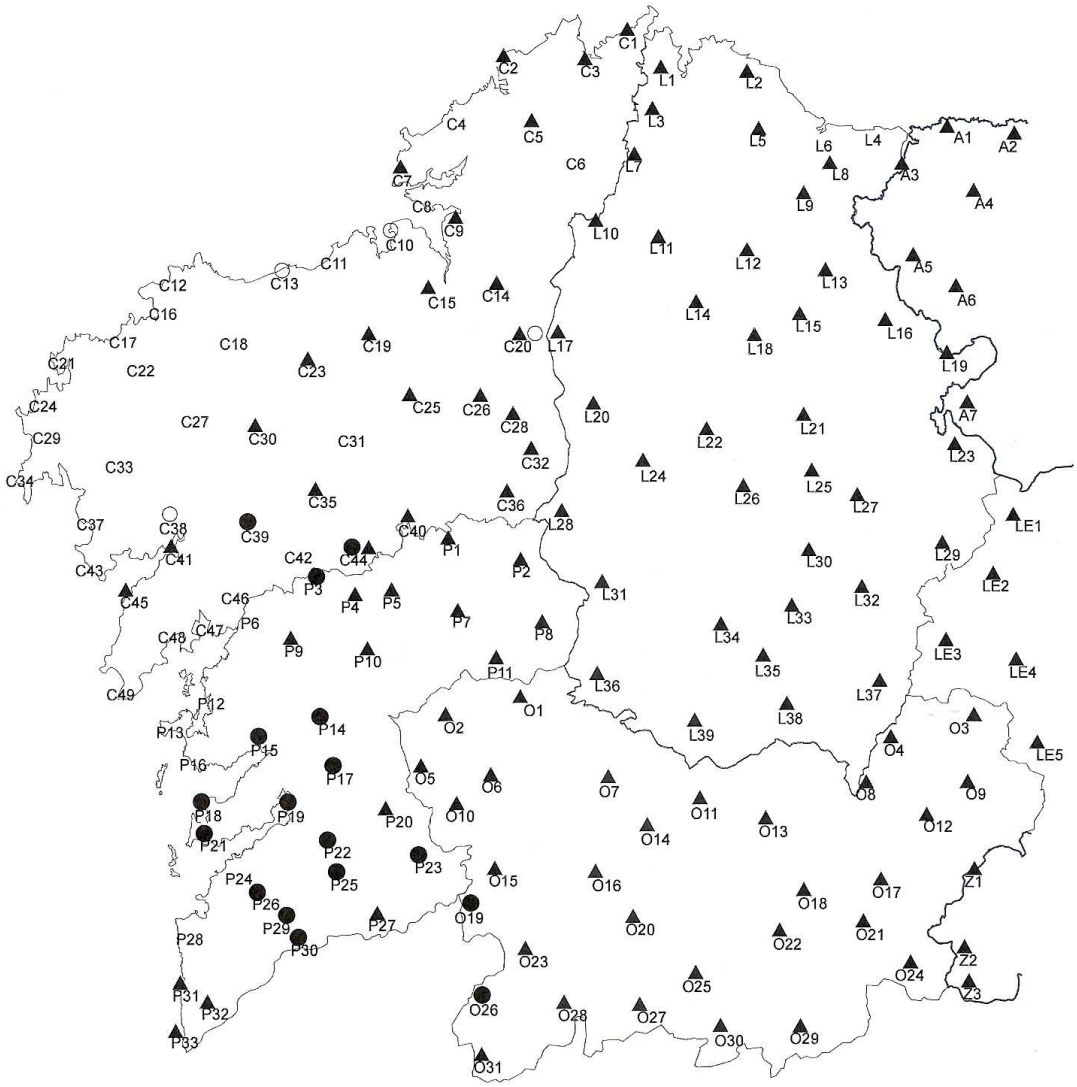

FIGURE 2. Map of “Magosto” ‘chestnut roasting’

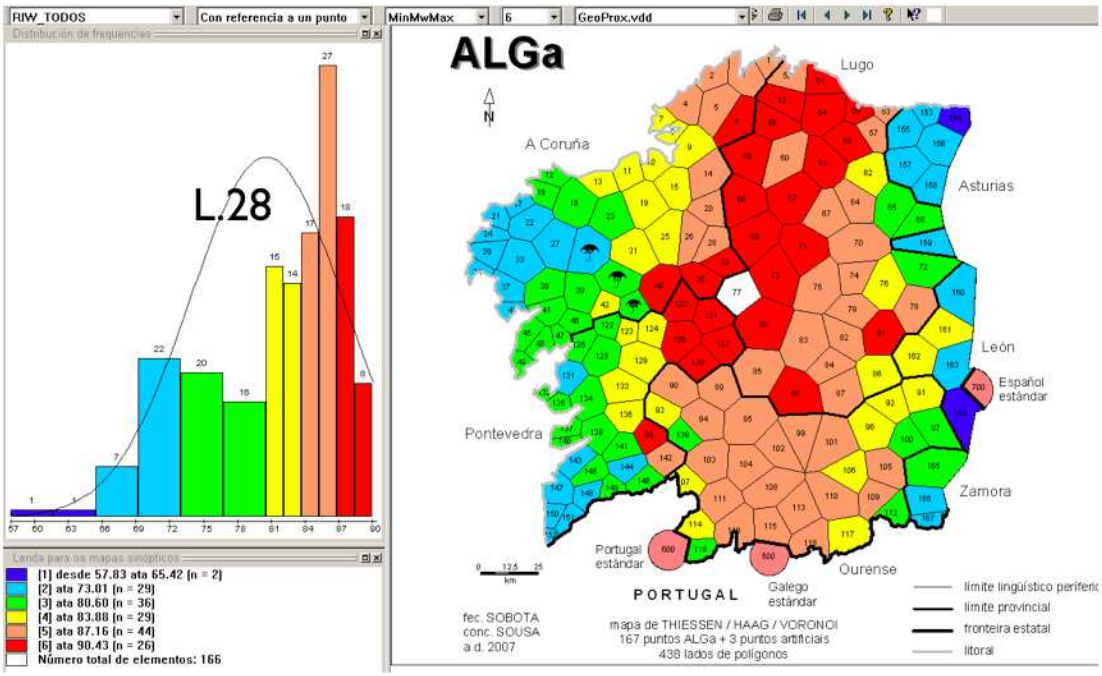

FIGURE 3. Similarity profile of L.28

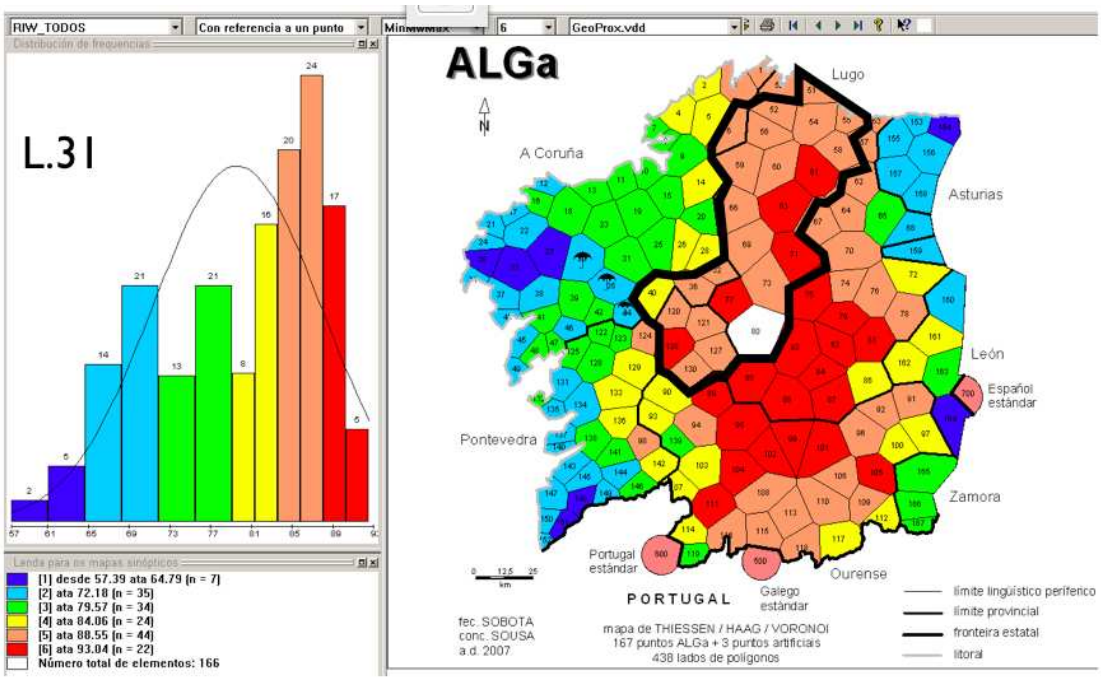

FIGURE 4. Similarity profile of L.31

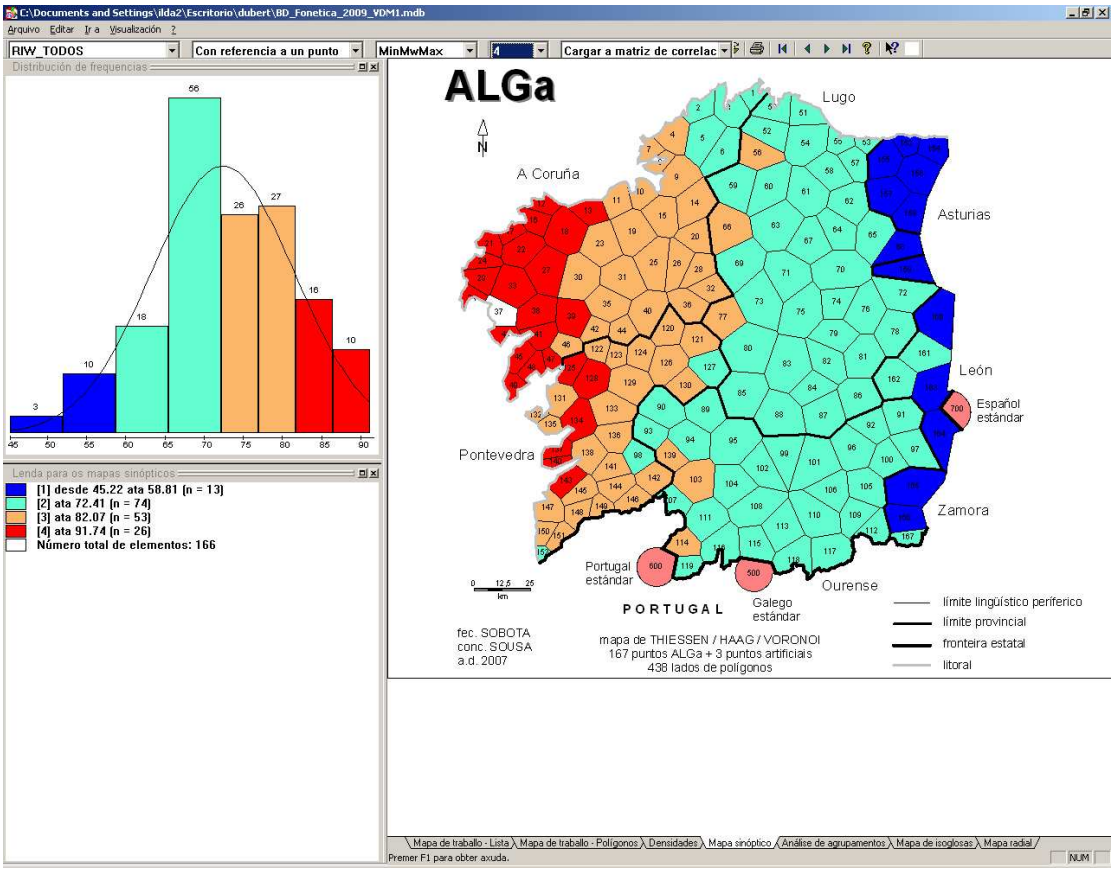

FIGURE 5. Similarity profile of C.37 with 4 cuts

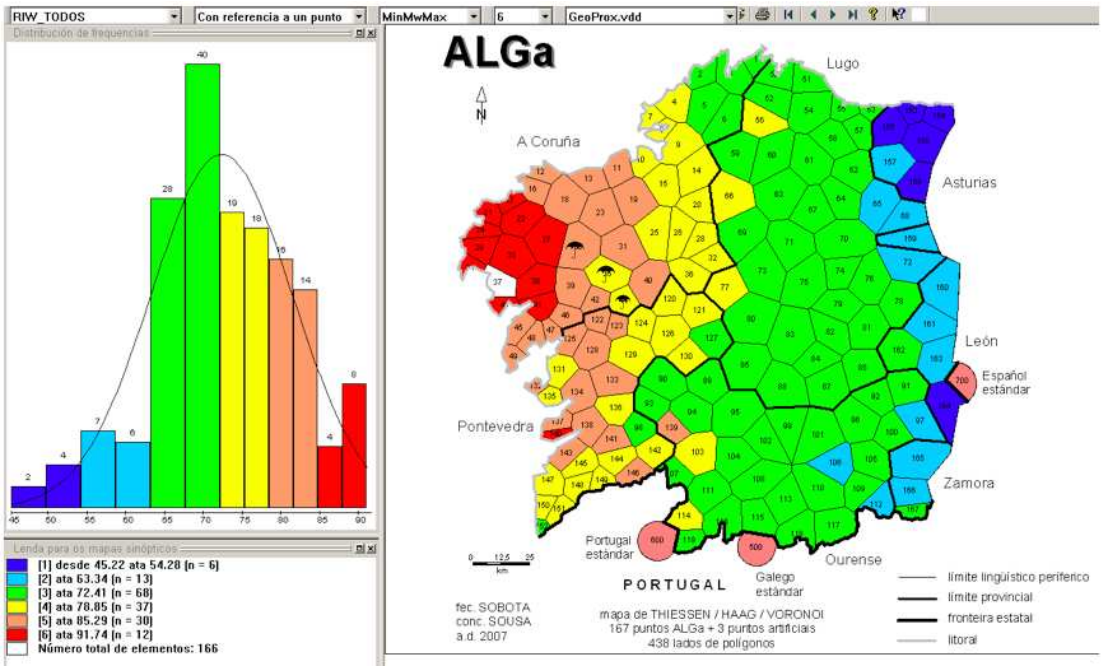

FIGURE 6. Similarity profile of C.37 with 6 cuts

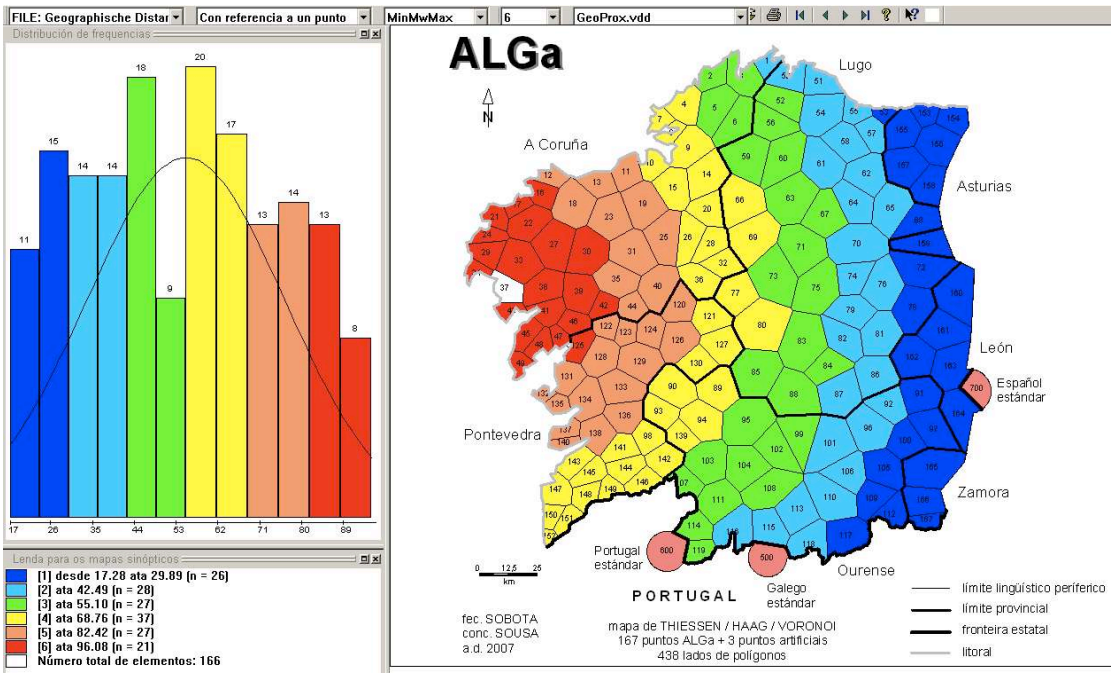

FIGURE 7. Proximity profile of C.37

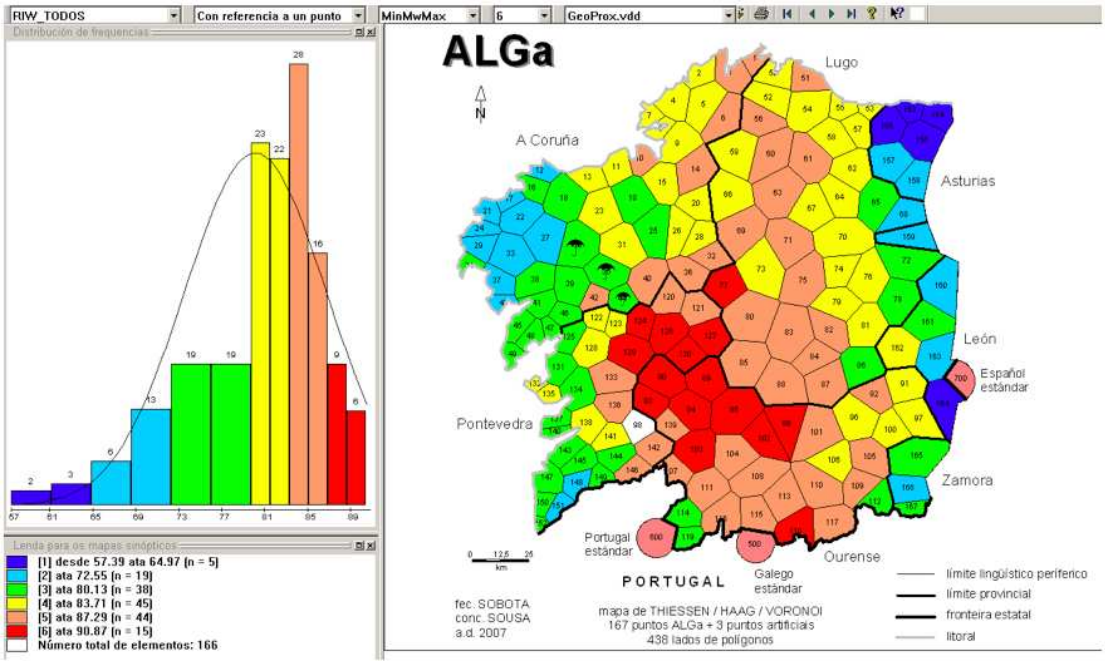

FIGURE 8. Similarity profile of O.10

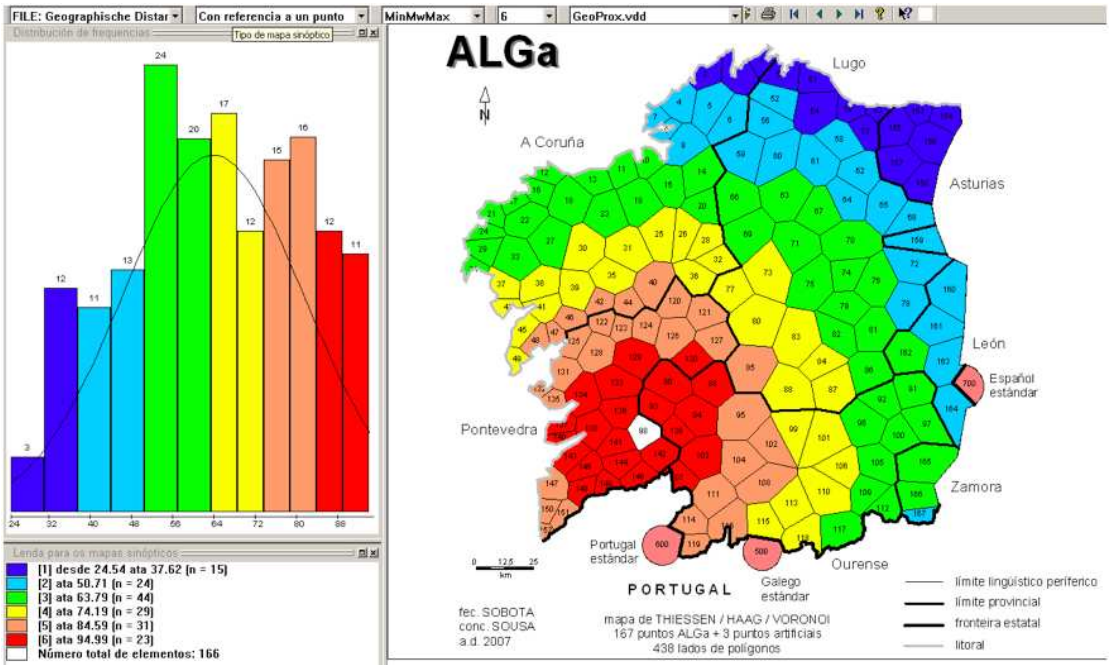

FIGURE 9. Proximity profile of O.10

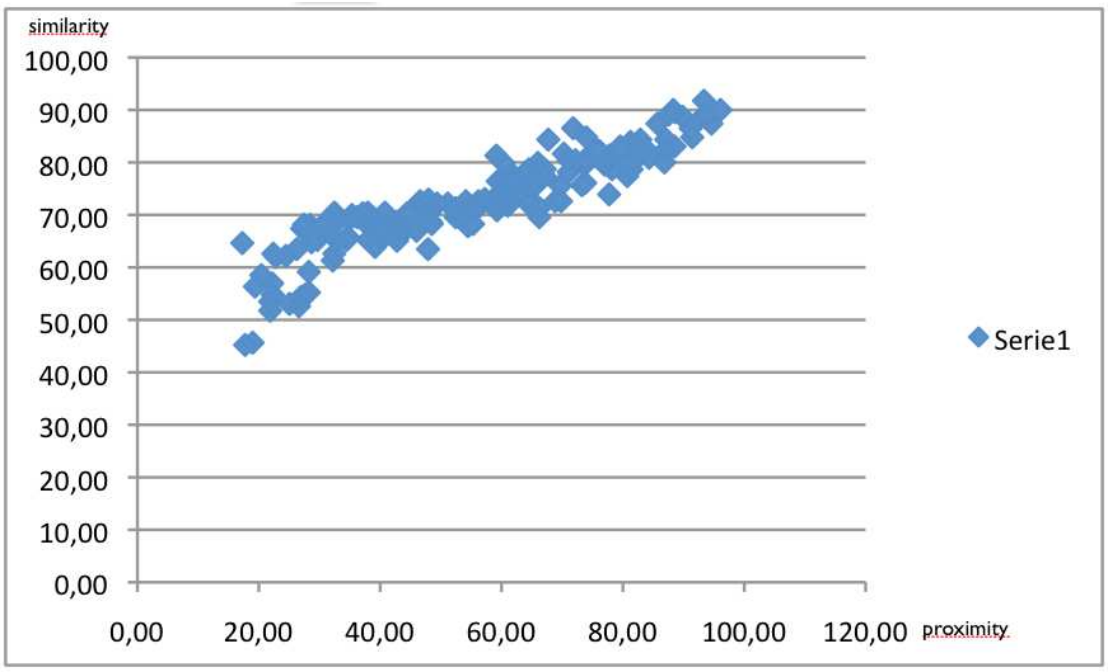

FIGURE 10. Scatter graph of the correlations between geographic proximity and linguistic similarity in C.37

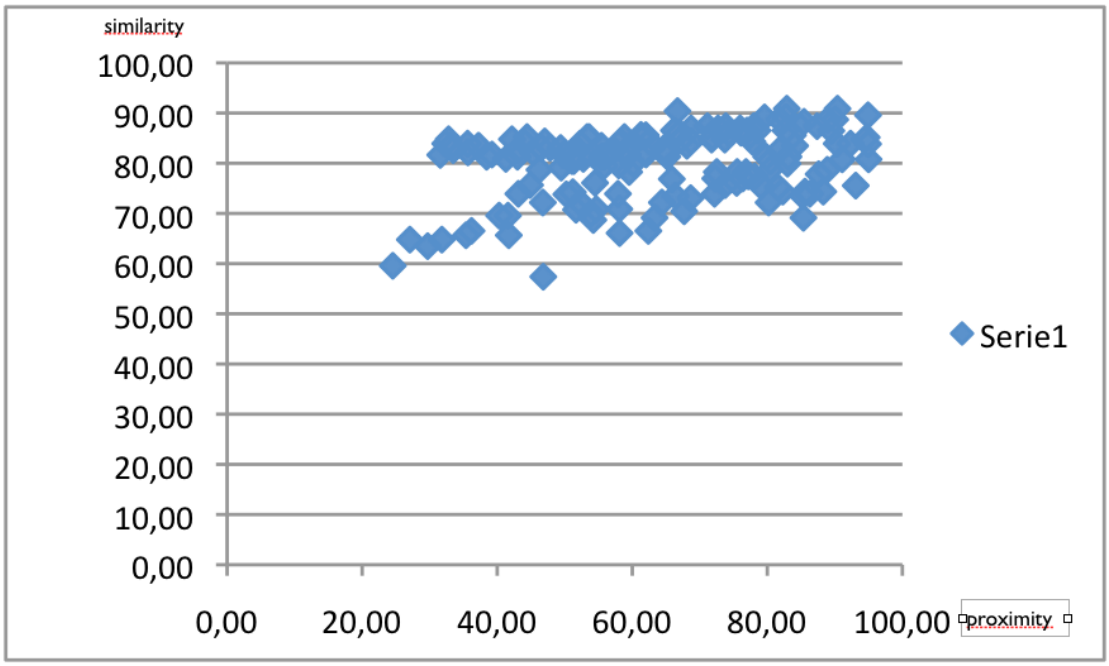

FIGURE 11. Scatter graph of the correlations between geographical proximity and linguistic similarity in O.10

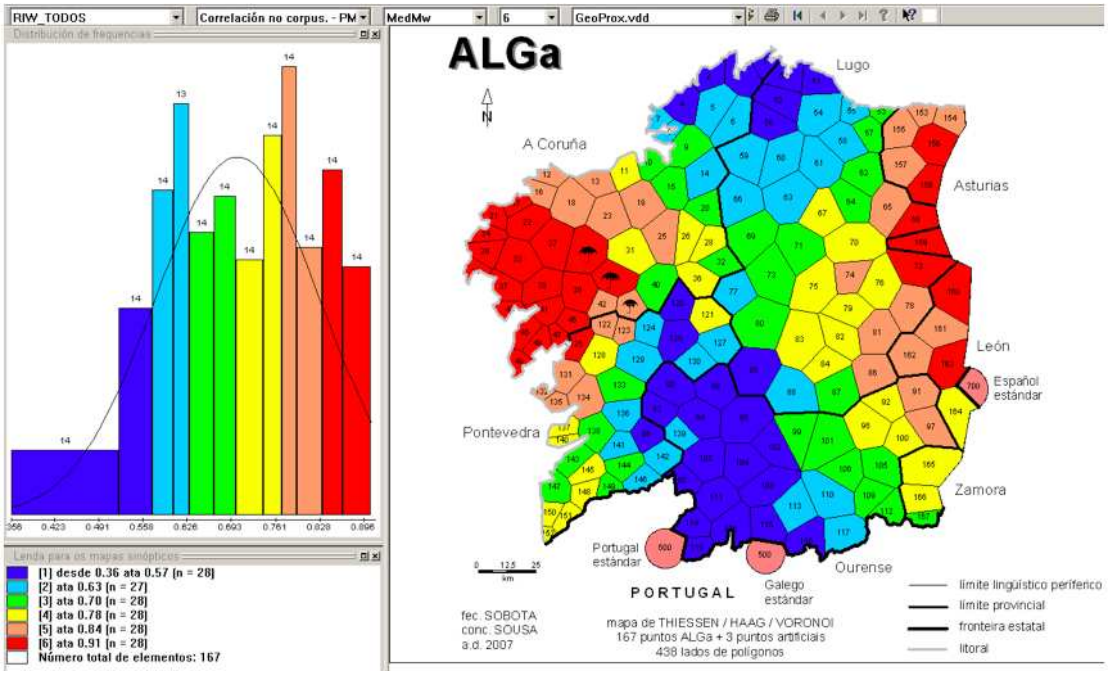

FIGURE 12. Choropleth of the geographical distribution of the correlations indices of proximity and similarity

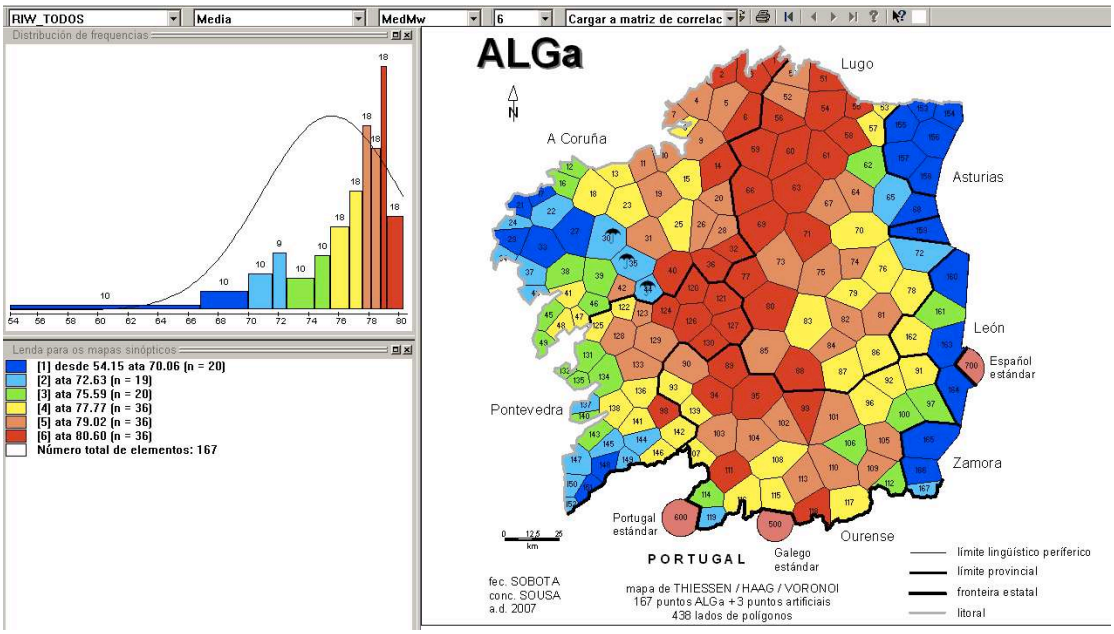

FIGURE 13. Choropleth showing the geographical distribution indices of the arithmetical means of similarity

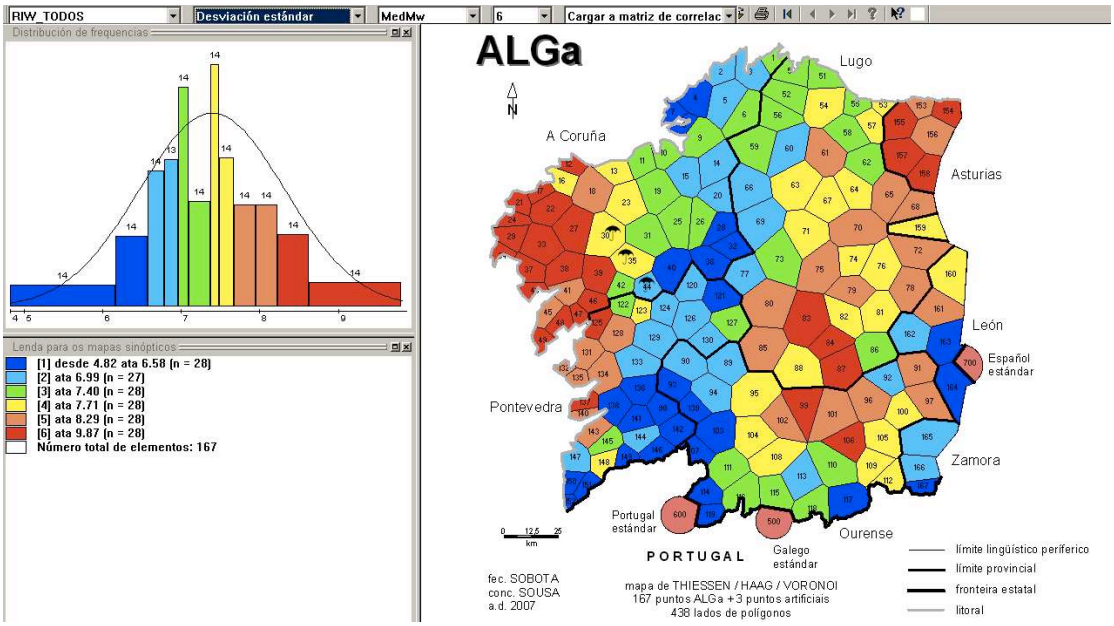

FIGURE 14. Choropleth showing the geographical distribution of standard deviation

POINTS	PROXIMITY	SIMILARITY
C43	96.08	90.00
C33	94.58	87.39
C29	93.53	88.70
C38	93.48	89.13
C34	93.35	91.74
C45	91.42	84.78
C41	91.16	86.52
C24	89.84	88.70
C27	88.62	89.57
C48	88.51	83.04
C22	88.30	90.00
C21	87.57	89.13
C47	87.49	82.61
C49	87.13	84.35
C39	86.84	82.17
C46	86.83	80.00
C17	85.69	87.39
C30	84.31	80.87
P6	83.70	82.61
C16	82.87	84.35
C42	82.73	81.30
P12	81.45	78.70
C18	81.24	83.91
P3	80.96	80.00
C35	80.73	77.39
P13	80.68	79.57
P9	79.55	83.04
C12	79.25	82.17
P16	78.18	78.70
C23	77.99	80.87
P4	77.76	81.30
C44	77.74	73.91
C31	77.00	79.57
P15	76.13	82.17
C13	75.83	82.17
P18	73.96	84.78
P5	73.79	76.09
P14	73.71	79.57
C40	73.66	80.44
P10	73.23	75.65
C19	72.18	80.44
P21	71.80	86.52
C25	71.07	78.17
C11	71.06	80.44
P19	70.27	81.66
P1	69.81	72.61
P17	69.67	75.55

POINTS	PROXIMITY	SIMILARITY
P7	68.83	72.61
P24	67.70	84.35
C26	67.13	77.39
C10	67.12	78.70
C15	66.27	78.26
O2	66.24	71.30
O5	66.22	69.57
P22	65.99	79.91
C36	65.98	76.09
P26	65.04	75.22
P2	64.57	75.22
P28	64.51	78.70
O11	64.39	72.17
P11	64.25	73.48
C28	63.70	77.83
P25	62.50	73.80
C20	62.09	77.39
C32	62.02	76.52
P29	61.58	75.65
C8	61.50	74.35
P30	61.05	73.04
P8	61.03	71.74
C9	60.60	77.29
C7	60.52	74.78
P20	60.31	79.91
L28	60.21	72.61
C14	60.11	75.65
P31	59.84	74.35
P23	59.37	76.42
O6	59.25	70.87
O1	59.17	71.74
P27	59.15	81.30
P32	58.80	73.04
L17	57.24	73.04
P33	56.44	71.74
L20	56.34	72.17
C4	56.20	72.61
L31	55.32	68.26
L36	54.48	67.83
O15	54.12	72.61
O19	53.91	70.74
L24	53.34	70.00
C5	52.82	70.44
L10	52.44	71.30
O7	52.44	69.57
C6	51.14	72.17
C2	49.36	72.17

POINTS	PROXIMITY	SIMILARITY
O16	48.50	68.26
L22	48.37	68.70
O23	48.20	70.74
L14	48.14	70.44
O26	47.97	72.93
L34	47.89	63.48
L11	47.24	72.17
L39	47.07	68.70
L7	46.58	72.61
L26	45.99	66.96
O14	45.88	67.39
C3	45.22	70.44
O31	44.43	70.31
L18	43.26	67.83
L35	42.93	65.65
O20	42.77	65.07
O11	42.77	67.39
L3	42.74	68.70
O28	42.20	66.38
L12	41.43	68.70
L33	40.99	67.83
C1	40.79	70.44
L30	40.10	66.09
L25	40.02	67.39
L38	39.40	64.35
L21	39.16	63.91
L1	38.79	69.57
O27	38.25	67.69
L15	38.22	69.13
O13	37.99	65.22
L5	37.96	70.44
O25	37.14	70.31
L9	36.19	69.57
L2	35.35	70.00
L32	34.90	69.13
L27	34.82	65.65

POINTS	PROXIMITY	SIMILARITY
L13	34.51	65.65
L37	34.47	65.22
O22	33.21	69.13
O8	32.79	65.22
L6	32.47	70.44
O18	32.45	62.61
L16	32.19	61.30
L8	31.66	66.52
O30	31.48	68.56
O4	30.66	68.26
L29	29.65	65.22
Le3	28.68	64.78
L4	28.42	68.12
A5	28.25	55.22
L23	28.24	59.13
L19	27.63	54.78
O17	27.46	67.39
O29	27.44	68.12
O21	27.14	67.39
A7	26.93	54.35
A3	26.62	52.61
O3	26.26	63.48
O12	26.25	63.48
A6	25.06	53.04
Le2	24.47	62.17
O24	22.81	62.01
O9	22.40	62.61
Le1	22.31	54.78
Le4	22.21	56.96
A4	21.89	51.74
A1	21.87	53.48
Z1	20.42	58.52
Z2	19.38	56.33
Le5	18.98	45.65
A2	17.75	45.22
Z3	17.28	64.63

TABLE 1. Two distributions of C.37. First column: names of points; second column: proximity indices; third column: similarity indices. The points are ordered by their proximity to C.37

POINTS	PROXIMITY	SIMILARITY
P22	94.99	80.79
O5	94.92	89.57
P23	94.92	83.84
P17	94.72	85.15
P20	94.36	84.72
P25	93.09	75.55
P27	92.33	83.91
P19	91.09	80.79
O6	90.39	90.87
P14	90.28	83.91
O2	90.05	88.70
O19	89.48	86.46
O15	88.95	88.70
P30	88.88	78.70
P26	88.29	74.35
P15	87.60	77.83
P10	87.37	87.39
P24	86.31	73.91
P18	85.57	74.35
O1	85.44	88.26
P29	85.35	69.13
P21	85.25	73.48
P11	84.75	87.39
P9	84.12	83.48
O23	83.79	85.59
P16	83.13	81.30
O26	82.99	79.91
P7	82.89	90.87
O7	82.78	89.13
O16	82.69	86.52
P12	82.29	74.35
P4	81.78	82.61
P5	81.78	88.70
P28	81.60	75.22
P13	80.46	80.44
P6	80.40	78.70
P32	80.20	72.17
P3	79.84	81.30
P8	79.59	89.13
O31	78.99	75.55
L36	78.55	86.52
P31	78.11	77.39
O14	78.00	87.39
O28	77.81	84.28
O20	77.37	85.59
C44	77.20	77.39
P2	76.99	86.52

POINTS	PROXIMITY	SIMILARITY
C47	76.85	78.26
C46	76.36	76.96
C42	76.10	86.09
P1	75.98	86.96
C48	75.53	78.26
P33	75.46	76.09
C40	74.29	86.09
L39	73.86	86.96
O11	73.86	87.39
C49	73.77	75.65
O27	73.71	84.72
L31	72.70	86.96
C39	72.52	78.26
C41	72.52	78.26
C45	72.29	76.96
C35	72.19	73.91
C36	71.86	85.22
O25	71.75	84.72
L28	71.12	87.39
L34	68.74	84.35
O13	68.71	86.96
C38	68.65	73.04
C31	68.09	83.48
C43	67.69	70.44
C32	67.28	84.78
L35	67.02	86.09
O22	66.90	84.35
O30	66.71	90.39
C30	66.37	73.48
L38	66.18	86.52
C25	65.86	76.86
C26	65.18	81.30
C28	65.05	82.17
L24	64.71	83.04
O18	64.61	82.17
C37	64.39	72.17
C27	63.30	69.13
L33	62.56	84.35
C33	62.38	66.52
O8	62.03	81.74
O29	62.01	85.59
C23	61.69	82.17
L20	61.29	85.65
L26	61.17	85.22
C19	60.15	80.00
O21	60.15	83.91
L37	59.54	78.26

POINTS	PROXIMITY	SIMILARITY
C20	59.29	81.74
L30	58.85	82.17
O17	58.84	85.22
L22	58.21	84.35
C29	58.13	66.09
C22	58.09	70.87
O4	58.00	83.91
C18	57.99	79.57
C34	57.88	73.91
C15	57.24	82.61
L17	56.69	82.17
L32	56.36	81.74
O24	55.86	79.48
O12	55.82	81.30
C13	55.50	81.30
L25	55.44	83.48
C24	54.67	70.87
C16	54.49	76.09
C17	54.23	68.70
C11	54.09	81.74
C21	53.84	70.00
C14	53.60	85.22
C10	53.32	85.22
O3	52.75	80.87
Le3	52.42	81.30
C12	52.32	72.17
L14	52.26	83.91
Z2	51.67	70.74
L27	51.49	82.17
L21	51.35	80.44
Z1	51.20	74.24
L18	51.08	81.74
O9	51.06	80.44
Z3	50.23	73.80
C9	49.83	81.66
L29	49.47	79.13

POINTS	PROXIMITY	SIMILARITY
L10	49.39	83.04
C8	47.84	83.04
L11	47.04	84.35
Le5	46.82	57.39
Le4	46.76	72.17
L15	46.52	83.04
Le2	46.26	78.70
C7	45.29	82.17
L23	44.87	75.65
L12	44.42	85.22
C6	44.14	83.91
L16	43.13	73.91
C4	43.04	82.61
C5	43.00	81.30
L7	42.18	84.78
A7	41.71	65.65
Le1	41.59	69.57
L13	41.28	80.87
L19	40.28	69.57
L9	39.23	81.74
L3	38.43	81.30
L5	37.27	83.48
C2	37.21	83.48
A5	36.20	66.52
L8	35.66	82.17
C3	35.61	83.91
A6	35.37	65.65
L6	33.49	82.61
L1	33.21	83.48
C1	32.80	84.78
L2	32.31	83.91
A3	31.80	64.78
L4	31.57	81.66
A4	29.70	63.48
A1	27.07	64.78
A2	24.54	59.57

TABLE 2. Two distributions of O.10. First column: names of points; second column: proximity indices; third column: similarity indices. The points are ordered by their proximity to O.10