DiG 19 (2011), 23-61 *DOI 10.1515/DIG.2011.002.*

Developing a database for dialectometric studies: The ALGa phonetic data. Dialectometrical analysis of 230 working maps[1]

Francisco Dubert

Abstract

In this article, I will present the problems of elaborating a phonetic database for a dialectometrical study. I will show how (1) the abstractness which is necessary to create the database conditions the interpretation of the results, (2) the difficulties of classifying the data as phonetic, morphological or lexical from a synchronic point of view, (3) the problems created by variation inside the same locality, a lack of data in some points and a situation of language contact (in this case, Galician and Spanish). Finally, I will present a dialectometrical study of the Galician dialects, made from the elaborated phonetic database. In this regard, I will present and comment on ten maps that allow us to show the inner structure of the phonetic data of Galician.

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 consisting of 321 working maps, 100 of which contained phonetic, 121 morphological and 100 lexical data. The study demonstrated that Galician dialects fall into two major areas, which can be divided into a western coastal strip and the remainder of the language's territory, respectively. The latter is in turn divided into a central and an eastern section (Álvarez, Dubert & Sousa 2006: 487–488). This classification suggests a greater degree of affiliation between the central

[1] This article (written within the project "Estudo dialectométrico das variedades dialectais do galego", financed by the Spanish Ministerio de Educación y Ciencia, HUM2006-06907) would not have been possible without the help of Marta Negro Romero, who drew up the database and worked with constant diligence, intelligence and a critical eye; her suggestions led to numerous improvements. I also wish to express my gratitude to Gotzon Aurrekoetxea, Ramón d'Andrés, João Saramago, Hans Goebl and an anonymous reviewer for their help in improving earlier versions of this article. All responsibility for any errors in this article is mine.

and eastern varieties of Galician, as opposed to the western varieties. This classification coincides roughly with that proposed by Fernández Rei (1990) using qualitative methods.

One part of the study (Álvarez, Dubert & Sousa 2006: 467–472) contained an analysis of 100 phonetic working maps used to produce three choropleth maps: two maxima of similarity maps (with four and six groupings respectively) and one showing asymmetry coefficients. The aforementioned study constitutes the direct antecedent of the present one, which is based on 230 phonetic working maps, more than twice as many as those used in the exploratory study that preceded it, and produced through a different procedure.

To present the results of this new database, §2 examines the quantification of data and the degree of abstraction underlying the database, and also evaluates the representativity of the sample. §3 takes a detailed look at the term *phonetic data*, describing the types of working maps underlying the database used to produce the maps. §4 studies the kinds of problems raised, when organising working maps, by divergent linguistic changes. §5 examines complications caused for the drawing up of maps by Galician-Castilian contact phenomena. §6 presents the results and §7 some brief conclusions.

2. Quantification and level of abstraction in the development of the database

Goebl, in one of his many papers discussing dialectometric methods (see Goebl 1992, 1993, 2002, 2003, 2006, and see 2008 for a comprehensive application to data for the whole Italian Peninsula), says: "Avant d'entamer une analyse dialectométrique, il faut être pleinement conscient et du *point de départ* et *du but de la recherche* à atteindre. En ce qui concerne le point de départ, il s'agit d'avoir une vision bien claire des avantages e des inconvénients de la collection atlantographique servant de base de donnés" (Goebl 1992: 432).

Given the strong and weak points of the collection of maps on which the data are based, it is as well to provide readers with detailed information concerning the characteristic features of this database to give them a clear understanding of what they are looking at when viewing the results. The reader should therefore bear in mind that "Tout comme l'atlas linguistique – qui, lui, n'est qu'une *représentation* approximative de la *réalité dialectal vivante* – la matrice des données ne constitue qu'une *image représentative* des données géolinguistiques répertoriées dans l'atlas" (Goebl 1992: 434).
Dialect atlases, which are our basic tool for creating a dialectometric database, constitute a simplified representation of reality, an abstraction produced by eliminating various types of information. The dialectometric database is a

second-order abstraction, a more refined representation, given that it is produced by eliminating even more data. It follows of necessity, then, that reality is always more complex than such a representation indicates.

Prior to the process of data collection for dialect atlases, it is necessary to select speakers who are considered representative and who meet certain conditions; this entails excluding other types of informant living in the place surveyed. The informants used provide responses that in many instances show a form of expression used by a larger or smaller proportion of the community in which they live but do not necessarily represent the community in its entirety, since not everyone speaks in the same way in any given community. Thus for example in Muros (in A Coruña province), some speakers use the form [esˈtaβ̞amos] 'we were (imperfect)' while others say [esˈtaβ̞anos]; and some say [ˈdeanos] 'we give (subjunctive)' while others have [ˈdeamos] or [deˈamos]. It is not always easy to include all responses on the maps because an informant may not always know, remember or even wish to recall all such forms.

Another factor that leads to simplification of the facts is the questionnaire itself. One cannot ask everything: dialectologists may seek a number of items that they consider representative of dialectally distributed phenomena, but the survey may bring to light unforeseen variations while other variations may fail to be collected. So for example, some Galician speakers say *m*[o]*to* 'motorbike' and others say *m*[ɔ]*to*; some say *c*[o]*ro* 'choir' and others say *c*[ɔ]*ro*; some say *f*[o]*to* 'photo' and others say *f*[ɔ]*to*; these forms are not covered by *ALGa.* It is also known that in the west of A Coruña province, where the phenomenon called *seseo laminal* occurs (see Vidal Figueiroa 1993), palatalization of syllable-final /s/ ([bes] 'you (sing) see' > [beʂ] ~ [beʃ]) is also widespread, yet such palatalization is scarcely covered in the *ALGa* data, perhaps because it only occurs within discourse and not in words pronounced in isolation. Neither is it possible to learn from the data in *ALGa* in what parts of Galicia dissimilation takes place between the copulative conjunction *e* (usually pronounced [ɛ]) and a determiner to its right, as in [j ɛʂtɪ] for /ɛ iste/ *e iste* 'and this' or [j ɔŋɐ] for /ɛ uŋa/ *e unha* 'and a', or vowel contractions of the type [ɛlɐ] /a ela/ *a ela* 'to her' etc.

Another issue is that *ALGa* has many localities where two, three or even four responses per question were given, whereas the dialectometric method requires that one response per place be chosen; the designer of the database is therefore called on to make decisions about which features to map and which to omit. This gives the impression that the excluded data do not characterize such places, so that certain similarities and differences are not registered and the following kind of situation ensues:

Locality A response *x*
Locality B response *x*

Locality C response *x*, *y*
Locality D response *y*
Locality E response *y*

If in C we choose *x*, C will be grouped with A and B; if we choose *y*, C will be grouped with D and E; either solution is a simplification of the real situation.[2] In the development of this database, the *marked* form was chosen in each case, i.e. minority responses or those which diverge most from the standard (thus preferring laminal [s] over apical [s̻], [ŋk] over [ŋg] in words like *ninguén* 'nobody', etc.).

ALGa also contains information about secondary points arising from simultaneous studies or given by informants who were present at the time of the survey. Since such information is not systematic, it is not incorporated into the dialectometric database. This again increases the abstractness of the data retained.

The present database consists of 230 working maps (the recommended minimum for studies of this kind being 100), distributed as follows (see Fernández Rei 1990 for an explanation of each phenomenon):

– 106 vowel maps: 70 on the quality of vowels in stressed syllables (covering metaphony, resolution of hiatus, and vowels followed by tautosyllabic nasals), 14 on the quality of stressed diphthongs (*ou ~ oi*, *oi ~ ui*, *ai ~ ei*), and 22 on the quality of unstressed vowels and diphthongs.
– 79 consonant maps (on metathesis, *gheada*, *seseo*, delateralization, etc.).
– 18 maps of interaction between consonants and vowels (such as *irmao ~ irmán* 'brother', *cheo ~ chen* 'full', *meo ~ medio* 'half').
– 11 maps on semivowels (such as *cando ~ cuando* 'when', *lambión ~ lambón* 'greedy').
– 13 maps about word prosody (*estar ~ tar*, *tose ~ tos*, *rede ~ re*).
– 3 maps concerning penultimate versus antepenultimate stress placement (in the words *paxaro*, *cantiga*, *parálise*).

The 230 working maps resulted in the establishment of 561 taxates or groups of variants considered as non-variant for statistical purposes. For example, for the working maps on voice in fricative consonants, all voiceless coronal fricatives in map 154 "Xunco" 'rush' of *ALGa III* were grouped together in a single taxate, regardless of their place of articulation, such

[2] João Saramago, reading an earlier draft of this article, communicated that in such cases he assigns a value of 0 to C in order to avoid making a subjective choice from the responses and thereby grouping C with either A/B or D/E. But my use of VDM does not allow for such a tactic and one value or the other has to be chosen. Note that Saramago works with distance maps, not similarity maps; in a similarity map C needs to be grouped either with A/B or D/E.

that 1 was assigned to voiceless lamino-postalveolar, apico-postalveolar and apico-alveolar fricatives, and 2 to voiced lamino-postalveolar fricatives. There are a mean of 2.4 taxates per map (contrasting with a mean of 6.9 found by Goebl in his 1992 study of northern Italy, giving some idea of the uniformity of Galician).

The VDM[3] programme performed 13861 inter-item similarity comparisons on these localities. So place 1 was compared to places 2, 3, 4 etc., place 2 (already compared to 1) was compared to 3, 4, 5 etc.; 3 (already compared to 1 and 2) to 4, 5, 6 etc.; 4 (already compared to 1, 2, 3) to 5, 6, 7 etc., and so forth. Each of these comparisons yielded a similarity index.

The number of working maps per phonetic variable depends on the data provided by *ALGa*; only a few questions found in the base remain unedited (such as *lagarto* 'lizard' in question number 1392, *testamento* 'will' in 2488, or *garlopa* 'jack plane' in 2651).

In cases where an *ALGa* question contains more than five gaps or non-cognate answers, no corresponding working map was produced. This accounts for the absence of maps for questions such as *pola ~ ponla* 'branch', *bicho ~ becho* 'bug', *croio ~ coio* 'pebble', *uce ~ urce* 'heather', *sarabia ~ saraiba* 'hail', which were present in the preliminary study (Álvarez, Dubert and Sousa 2006). The reason is that each blank response in *VDM* counts as a "difference", and such differences would imply a false distribution: in a map such as *bicho ~ becho* 'bug', for instance, blank responses would count as if there were some such thing as **bacho* or **bocho*, whereas there is actually no cognate response. The more blank spaces are admitted, the more the information in the database will be distorted as a result.[4]

In this respect it is important to observe that the indices contained at the locations C.30(*30*),[5] C.35(*35*) and C.44(*44*) are less indicative and reliable than those from elsewhere; in these places a large number of questions were left blank because the questionnaire was not completed. If we had disposed

---

[3] VDM (*Visual DialectoMetry*) is a computer programme created by Edgar Haimerl in Salzburg between 1997 and 2000. Using data taken from a database, it permits automatic calculations of indices of similarity, skewness, etc, and the creation of different kinds of dialectometrical maps. Basic information in http://www.dialectometry.com/dmdocs/englisch_fr.html (16th July 2010); see also Rivadeneira and Casassas (2009).

[4] João Saramago has pointed out (personal communication) that, without using *VDM*, in such cases one may omit the locality and calculate the indices with one locality fewer, so that although some distortion is produced in the data, the lack of responses will not be counted as distinct responses.

[5] The references of localities in *ALGa* differ from those given in these maps. I will identify localities by means of their *ALGa* reference, in Roman, followed by their number on the map (in parentheses, in italics). For example, locality L.5 in *ALGa* corresponds to item 54 on the map, and will always be referred to here as L.5(*54*), and so on.

of the missing responses, these three adjacent localities would have different indices. Goebl (1981, 1992) marks places in northern Italy with deficient responses with an asterisk in order to remind the reader to treat the information with caution or even to leave it out of the dialectometric analysis ("Il est bon de les exclure de l'interprétation synoptique des cartes choroplèthes", Goebl 1981: 369).

A word of caution is also necessary regarding the actual representativeness of the sample, given that processes represented on different maps are not equally general in scope. While maps such as those for *gheada*, coronal fricatives, or stressed vowels in words such as *serra* 'saw', *neto* 'grandson', *mirar* 'to look' have broad application across the lexicon (many words contain such segments or comparable patterns), other maps refer to smaller numbers of words. Such is the case with maps for metaphony of unstressed final [e] over stressed [ɛ] in words like *sete* 'seven' (*ALGa II*, map 82), *dez* 'ten' (*ALGa II*, map 285), etc. Here, since each word has its individual history and each result has a specific geographical distribution, it was decided to include several maps. Consequently, five maps represent this type of metaphony. That is a point to be borne in mind when evaluating the results presented here. *Gheada* and *seseo*, which are phenomena that affect very many words, are only represented by six maps each, oriented to three different aspects of each of these phenomena. Clearly, then, [ɛ]...[e] metaphony is over-represented, and this fact results in over-weighting the southwestern areas in which that phenomenon occurs.

Finally, the reader should recall that all the maps in the database contain phonetic variation. No map contains a phonetically identical response given everywhere; at the very least, there will be a phonetically differentiated response in some locality or other for each question. Maps showing a single response were not included, because the aim of dialectology is to study variation; indeed, maps with a single response are rare among the material published by *ALGa*. Thus in *ALGa III* there are only 33 maps with a single response, some of which contain no new information, such as those showing the reflex of initial PL- (*chegar* 'to arrive', *chorar* 'to cry' and *chuvia* 'rain'). Therefore, the indices of similarity given in the maps are lowered: real similarities between the dialects reflected by *ALGa* are greater than those reflected in the database.

3. What is meant by *phonetic* data? Phonetic, morphological and lexical data. Synchrony and diachrony

Methodological issues may be compounded by certain theoretical questions. Since the maps presented in this study refer to phonetic data in *ALGa*, I shall

begin by considering the differences between phonetic, morphological and lexical processes. Although long a part of traditional dialectology (witness the way materials are classified in *ALGa*'s own publications), these categories merit more detailed study than falls within our present scope.

In Iberian dialectology it is usual, by and large, to classify relevant dialect phenomena as *phonetic*, *morphological* and *lexical*. Yet such a classification, although neither explicitly acknowledged nor questioned, is nevertheless problematic, since it is not always a straightforward matter to determine whether a given variant *x* should be regarded as a case of category *Y* or category *Z*. Indeed, the authors of the preliminary study recognise the difficulty of differentiating between so-called phonetic and morphological phenomena:

> The 100 questions in the area of phonetics were chosen bearing in mind the classical categories of historical phonetics and certain traditional tendencies in Galician phonetic studies. The 121 questions about morphology cover both nominal and verbal morphology. In this context it is not easy to draw the line between phonetics and morphology, but responses were classified taking into consideration variations that affect noun and verb forms and are customarily regarded as morphological in nature. (Álvarez, Dubert & Sousa 2006: 464; my translation from Galician)

To appreciate the consequences of such difficulties, a cursory look at the published volumes of *ALGa* will suffice:

a) In *ALGa III*, we find maps number 29 "Sogro" 'father in law', 30 "Novo" 'new' and 31 "Corpo" 'body', subtitled "Reflexes of stressed Ŏ in words ending in -U", which show the effect of final-vowel metaphony on the vowel of a stressed syllable. In *ALGa II*, we encounter the same data in maps 4 "Sogro, sogra", 5 "Corpo" and 7 "Novo, nova". The reason for their inclusion in *ALGa II* is not *phonetic*, but rather the fact that in *s*[o]*gro* ~ *s*[ɔ]*gra* or *n*[o]*vo* ~ *n*[ɔ]*va* (as opposed to *s*[ɔ]*gro* ~ *s*[ɔ]*gra* or *n*[ɔ]*vo* ~ *n*[ɔ]*va*) the root-vowel alternation is considered to be morphologized (compare *s*[o]*gro* to *c*[ɔ]*rvo* or [ɔ]*vo*) in such a way that gender correlates with stem-vowel alternation: thus the masc. has /o/ as both root and stem vowel, while the fem. has [ɔ] as the root vowel and [a] as the stem vowel.

b) In *ALGa III*, we find 79 "Colliches" 'you took (perfect)' representing the reflex of a stressed ĭ in words ending in -ī (/e/.../i/ metaphony), with the variants *colliches* and *colleches* being shown. But *colliches* is a paradigm form of the verb *coller* 'to take' no different from all regular second-person singular past indicatives of the second conjugation (like *comiches* 'you ate', *bebiches* 'you drank', etc.): this *colliches* stands in contrast to *colleu* 'he took (perfect)' and *collemos* 'we took (perfect)', and likewise

to *colles* ‘you take (sing)’, and *colleras* ‘you had taken’. It therefore also appears in *ALGa I* in map 51 “Colliches (VT)”.

c) In *ALGa III*, we find map 77 “Lingua” representing stressed ĭ in a syllable that ends in a nasal; the variants *l*[i]*ngua*, *l*[e]*ngua* and *l*[ɛ]*ngua* ‘tongue’ are shown. The same data reappear together with other *phonetic* data in map 19 “Lingua” in *ALGa V*, which shows cognates of LĬNGUA (in contrast to map 27 “Úvula”, which gives the variants *galillo*, *campaíña*, *pingallo*, etc.). But *lingua* ‘tongue’ belongs to the semantic field of body parts: it is thus related to *cabeza* ‘head’, *ollo* ‘eye’, *boca* ‘mouth’ etc. At the same time, the lexical map of *lingua* needs to show all the forms *l*[i]*ngua*, *l*[e]*ngua* and *l*[ɛ]*ngua* (not to mention variations between *ng* and *nc*), since these all constitute phonological representations of the lexical entry in a given dialect; no rule can predict either the height of the vowel nor the voicedness of the velar stop.

This last example suggests two further considerations. One of these concerns the opposition between *synchrony* and *diachrony*, while the other has to do with the synchronic *productivity* of phenomena shown on the maps.

Historically, the emergence of linguistic geography is inseparably linked to interest in ascertaining whether the neogrammarians’ views on the regularity of phonetic change were correct. Linguistic geography ended up also providing valuable linguistic and extralinguistic material that contributes to explaining change, but also new data demanding explanation. As is observed in the preface to *ALGa III*:

> One of the first decisions to be made when planning a set of maps for the purpose of studying the phonetics of a language concerns the angle from which phenomena are going to be studied. It must be determined whether the approach will be synchronic or diachronic, and it should be clarified whether a phonetic or a phonological point of view will be adopted. (*ALGa III*: 11; my translation from Galician)

It is well to caution readers about this because the same thing that is viewed in the diachronic dimension as a phonetic change often corresponds to a morphological alternation in synchronic terms. Consider e.g.

TENEO > *teño* ‘I have’
TENES > *tes*, *teis* ‘you (sg.) have’
TENET > *ten* ‘he / she / it has’
TENEMUS > *temos* ‘we have’
TENETIS > *tendes*, *tedes*, *teis* ‘you (pl.) have’
TENENT> *ten*, *tein*, *teñen* ‘they have’
TENEBAM > *tiña*, *tía* ‘I had (imperfect)’

The allomorphic variation in the present indicative of the verb *ter* ‘have’ ceases to be a regular reflex of phonological or phonetic phenomena: no

regular synchronic phonological process accounts for the presence of the nasal in *teño ~ teña* or its absence in *tes ~ teis*.

The preface of *ALGa III* goes on to state that the main interest of the phonetic information brought together in this volume is diachronic: it serves to show how a given phonetic sequence (e.g. -ANUM) developed in a certain geographical area and the various cognates that ensued (*-an*, *-ão*, *-ao*, *-au*, *-á*). This diachronic orientation of *ALGa III* determines the one adopted in this study: if *ALGa* is the source for the data in this dialectometric analysis, the analysis can only reflect *ALGa*'s diachronic articulation. But the following must also be borne in mind:

a) The transcription and phonetic description of *ALGa*'s data is partly *conventional*. Detailed studies currently underway of isolated phonetic processes (such as Martínez-Celdrán and Regueira 2008, Labraña Barrero 2009) demonstrate that real phonetic phenomena are less regular than this.

b) Synchronic alternations of an allophonic and automatic kind that do not affect distinctive features (Mohanan 1995) are not normally the object of treatment in *ALGa*. Consider the treatment of nasal consonants in *ALGa III:* there are 54 maps showing reflexes of Latin intervocalic -N- (*e.g. irmán*, *chan*, *veciño*, *veciña*, *unha*), two concerned with the emergence of nasal consonants from nasal vowels (of the type seen in *miña* 'mine') and three with the appearance of a non-etymological nasal vowel (as in *ponla* 'branch'); but not a single map is concerned with the synchronic treatment of syllable-final nasal consonants. The most *productive* component of the allophonic phonology of present-day Galician is virtually absent from *ALGa* and is merely represented in a conventional manner in the phonetic transcriptions.

On the other hand, phonemic variation involving distinctive features in strictly phonological contexts, such as e.g. rhotacism as seen in [oɾ mãŋkoʂ̪] "os mancos" 'the one-handeds' (vs. [oʂ̪ takoʂ̪] "os tacos" 'the plugs'), is well represented. So too are differences in segment inventories such as [ɦato] vs. [xato] vs. [gato] "gato" 'cat'). Also well represented are segments possessing features that are not necessarily distinctive yet are prominent and easily perceived, even for the non-specialist. These are the features referred to by Labov (1972) as *indices* and *markers*. Trubetzkoy (1939: 46–48) had already written about these linguistically non-functional features that nonetheless have social and stylistical relevance. This is the reason why all types of coronal fricatives (*apico-alveolars*, *dento-alveolars*,

lamino-alveolars, *apico-postalveolars*, *lamino-postalveolars*, etc.[6]) all appear explicitly in *ALGa*.

However, many of the phonological variations covered in *ALGa* are lexicalized. For example, in western A Coruña province, [ɔ] is used in words such as *flor* 'flower', *maior* 'greater', *calor* 'heat', *olor* 'smell', but [o] in words like *mellor* 'best', *peor* 'worse', *corredor* 'corridor' or *motor* 'engine'. Whatever future studies may have to say about this, at present the quality of these stressed vowels cannot be linked to a particular morphological marker or phonological context: it is simply the case that some words show [ɔ] while others have [o], resulting in quasi-distinctive pairs. The same applies to contrasts such as *voz* 'voice' and *sol* 'sun', which are pronounced with [ɔ] and [o] in some areas. In the present-day synchronic system, these vowels are specified in the lexicon, just as the vowels are in *piso* 'flat', *peso* 'weight' and *paso* 'step', so strictly speaking there is no justification for talking about any *phonetic* phenomena here.

The same may be said of prosodic variation. In northeastern Lugo province and in Asturias, a form *largato* 'lizard' is found which is also present in the Asturian language area. At some point in its history, this form arose by metathesis from *lagarto* (possibly by analogy with *largo* 'long'). Now synchronically, in the present state of these dialects, there is no place-changing of underlying /r/ phonemes, and no context which gives rise to such an alternation productively. All we have here, then, is the outcome of a phonological process that affected one lexical item.

Given that these forms are cognates of the same Latin etymon, I opted to include them in the database since they represent interaction between phonological changes and geography. For the most part, however, these rules pertain to the historical grammar of each dialect.

4. Problems arising directly from geographical variation

Another issue that has to be dealt with when producing working maps is the problem presented by geographical variation itself. This has the effect in some places of preventing the creation of the context that is necessary for the operation of a given rule or for the emergence of a certain distribution.

---

[6] Note that these indices function as factors of geographical differentiation. The laminal / apical opposition is not generally considered distinctive; the lamino-alveolar and apico-alveolar consonants are assumed to stand in opposition to the lamino-postalveolar for the passive articulator. In dialect A, the apico-alveolar and the lamino-postalveolar are in an alveolar / postalveolar opposition; in dialect B, the lamino-alveolar and the lamino-postalveolar are in an alveolar / postalveolar opposition. The apical or laminal features are what distinguish A and B.

Thus when we wish to chart a phenomenon in Galician dialectology that is as important, qualitatively and perceptively, as that of the contraction of vowels in hiatus of the type -[ɔa] > -[ɔ] – which is characteristic of a well-defined area of southwestern Galician with continuity in Portugal, yielding a variable according to which two areas may be distinguished, in one of which the contracted form *mo* 'tooth' occurs while in the other the uncontracted form *moa* is encountered (*ALGa III*, map 124) –, it turns out that since Latin intervocalic -L- was not dropped in western Asturias, the context in which this contraction might or might not have occurred is absent from that Galician dialect (Galician [ɔa] ~ [ɔ] comes from Latin -ŎLA, and the Asturian dialect of Galician preserves the form *mola*). Now if, in the working map for *moa* ~ *mo*, we incorporate the information that Asturias has *mola* and southwestern Galician has *mo* while *moa* occurs elsewhere, the presence of intervocalic /l/ in the Galician of Asturias will be over-represented given that there were already other maps focusing specifically on the distribution of loss or retention of intervocalic -L-.

	Western Asturias	Galicia	
Loss of intervocalic -L-	*mola*	*moa*	
		Southwest	Elsewhere
Contraction of -[ɔa] > -[ɔ]	*mola*	*mo*	*moa*

Comparable difficulties also present themselves in other maps, and strategies for dealing with these had to be chosen on a case-by-case basis. In the case of *moa*, it was decided to assign 1 to *mo* and 2 to everything else (i.e. *moa* or *mola*).[7] The working maps for *quente* 'hot' and *xear* 'freeze' cover the issue of the presence of -[l]- in Asturias, *calente* and *xelar*, respectively.

The map for *xunco* 'rush' (*ALGa III*, map 154), which serves as the working map on the presence of syllable-initial voiced coronal fricatives, passed over the existence in some dialects of lamino-alveolar as opposed to post-alveolar and apico-alveolar fricatives. The latter information is covered in other working maps produced from the data for *xenro* 'son in law' (*ALGa III*, map 154) and *xaneiro* 'January' (*ALGa III*, map 151), in which, conversely, information on the voice feature of the fricatives was omitted. Rather than compiling a single map containing all this information, in which the

[7] The operation of grouping variants and assigning them numbers is called *taxatation.* With this process, we reduce different answers to one, known as a *taxate*, which is assumed to represent a common feature we wish to represent on a map. Taxatation is an important process because it determines the basic cut-off points when producing maps. Geolinguistic variation poses major difficulties for taxatation.

phenomena mentioned would interact, it was considered preferable to produce several maps:

a) maps for *casa* 'house' (*ALGa III*, map 196) and *xunco* 'rush' were used to register values for the voice feature, assigning 1 to [-voice] and 2 to [+voice];
b) maps for *xaneiro* and *xenro* were used to register variation between [ʃ] ~ [s̟] ~ [s] ~ [s̠], assigning 1 to [ʃ], 2 to [s̟], and so on.

According to Goebl, "il faut donc distinguer les différentes cartes *originales* [of the atlases, FDG] et les cartes *de travail* de la matrice de données", because "d'une seule carte d'atlas, il est possible de tirer, par la taxatation multiple de mêmes données suivant des critères métroloquiques différents, plus d'une carte de travail" (Goebl 1992: 436).

In such cases, it may be said that we are looking not so much at maps of words as at maps of phenomena. This is a different kind of map from that created in the preliminary study (Álvarez, Dubert and Sousa 2006), where the map for *veciño* 'neighbour' was given a taxate for each phenomenon: 1 for dental fricative, 2 for apical fricative, 3 for laminal fricative, 4 for voiced fricative, so that distinct phenomena were grouped together in a single map. In the present version of the database, in contrast, there are separate maps for *seseo* and for voiced coronal fricatives, for instance.

The existence of distinct lexical types as responses to a given question can also give rise to more than five gaps, making it impossible to obtain maps with meaningful information: thus no working map for *ALGa III*, map 4 "Cántiga" 'song' was obtained in which to record variation in stress-placement in *cántiga* vs. *cantíga*, because other responses were obtained in twenty places. Likewise, working maps could not be produced, in spite of their potential interest, for *ALGa III*, map 12 "Magosto" (with variation between *magosto* and *magusto*) or *ALGa III*, map 69 (with the variants *bicho* and *becho*), because there were too many places where no cognate was collected.

5. Language contact situation and interference

The tendency for traditional Galician solutions to be displaced by Castilianisms is another source of complications. A Castilian-influenced lexical item may entail the importation of Castilian phonology, as when the word *colegio* 'school' brings with it the /x/ phoneme, or lead to the loss of native forms that would have contributed information about phonological variation, as when the intrusion of Spanish *ganado* 'cattle' makes it impossible to know whether, in a particular locality, the traditional form that existed

previously had been *gando* (with epenthesis of a nasal consonant: *gãado* > *gando*) or *gado* (with complete loss of the nasal feature: *gãado* > *gado*).

We have two options here: one may adopt either a purist approach to Castilianisms or a realistic one, and each entails certain costs and benefits. One of the costs of the purist approach of concealing or avoiding Castilianisms is that it falsifies the real situation, for whether we like it or not loans from Castilian do form part of the present lexical stock of Galician: to conceal them is to conceal part of the real state of the language. Another is that ignoring Castilianisms results in the impossibility of including maps for the many gaps in the database that ensue from this policy. As we have seen, there is no map in the database that lacks responses for more than five localities. Discounting Castilianisms, however, will multiply the number of blanks.

A benefit of purism is the avoidance of noise and distortion. Recall that one of the purposes of a language atlas is, precisely, to collect valuable information before it succumbs to the spread of innovations radiating from the official language variety. Up to a point, this purpose is often associated with *purist* attitudes involving preference for a *traditional* response over the standardizing innovation, the latter tending to be perceived as a deformation of an earlier more pristine stage. In the case of Galicia the official language variety that has caused modifications to the traditional forms of speech is standard Castilian (the data being used are from 1975–1977, before Galician became standardized and the language was given official status).

In the development of this database, I adopted a mildly purist stance[8]. Some maps were omitted because, owing to lexical Castilianisms, more than five gaps would have occurred in the data. For example, the aim of maps such as 12 "Coitelo" 'knife' or 13 "Martelo" 'hammer' in *ALGa II* is to map metaphony of the type /ɛ/.../o/ (e.g. *mart*[ɛ]*lo* vs. *mart*[e]*lo*). The present of Castilian loans like *martillo* or *cuchillo* tells us nothing about Galician metaphony. Admitting such data would have meant distorting the purpose of the working map.

Nonetheless, some maps containing Castilianisms were included because of their value in defining areas. Such is the case of a map such as that for *lingua*, in which the stressed [ɛ] that shows up all over the Galician

[8] I acknowledge that expressions like *mildly purist* are problematic. When is someone being *mildly* purist? What constitutes being *too* purist? What determines the limits? Why should *siempre* 'always' be allowed into the database, but not *cuchillo* 'knife'? Is it possible to achieve full consistency? It is unfortunately not feasible to explain in detail here the decision made in each and every case, just as it is also doubtless impossible to provide explanations that would satisfy everybody. Such decisions must also be made when various different responses occur in a given locality and only one them can be chosen.

language area in forms such as *lengua* may very well be a Castilian loan in origin, yet when we examine this map two clear blocks with *lingua* (with [i]) emerge, one in the Costa da Morte area and the other in Asturias, a fact that can hardly be ignored. Again, in the map for *sempre* 'always', the Castilian form *siempre* shows up all over eastern Galician, but there are also congruent areas with the authentic Galician stressed [e] in the south. Given this, I created working maps based on the *ALGa* maps in question.

An interesting task that remains to be done using the dialectometric method presented here would be to analyse a substantial number of lexical maps that include Castilian lexical loans in order to find out whether certain parts of Galicia are more heavily castilianized than others.

6. The dialectometric analysis

Granted all the above provisos, let us now turn to the maps. I shall discuss the results of applying the dialectometric method to the phonetic data of *ALGa.* First, I shall present some comparisons between localities. Next, I will discuss a map showing maxima of similarity, followed by an analysis based on mean similarities. This is complemented by an analysis of standard deviation. In the next, step I will present an analysis of asymmetry coefficients and an isogloss analysis, before concluding with an analysis of dendrographic grouping.

6.1. *Similarities between localities (see figs. 1–4)*

As explained above, the VDM programme compares each locality in the database with each and every other locality, and in this way similarity values are obtained between all the places, giving a total of 13861 such values[9]. This number of values permits us to perform several dialectometric operations. Obviously the simplest of these is simply to compare a given locality with the totality of others in the data network. Fig. 1, with four-

[9] The indices were produced by applaying the ponderation algorithm RIV_{lk} (Relative Identity Value) to the matrix of data; this algorithm contrasts with the $WIV(1)_{lk}$ (Weighted Identity Value). The former is used when the dialectologist wishes to give the same statistical status to all the dialectal features; the latter when the dialectologist wants to highlight rare, less common features (Goebl 1987: 67–79). The reason for choosing RIV_{lk} is the relatively small range of Galician and the low rate of taxates per working map, namely 2.4.

colour scales[10], shows the two localities with the lowest mutual similarity index (this is the reason why I have selected C.29(*29*) to generate the map): in the west, C.29(*29*), and in the east, Le.5(*164*), a locality where a variety of Leonese is spoken. These have 41.74% shared responses.

As can be gleaned from the legend, there are 12 very dark grey[11] points with similarity indices ranging between 41.74% for Le.5(*164*) and 53.91% for L.23(*72*). On the other hand, the locality with the highest similarity with C.29(*29*) is C.27(*27*), their similarity index being 92.17%.

Maps like this are the best way to show geolectal variation. Light grey areas represent similarity indices between 79.87% and 92.17% and are seen to surround C.29(*29*). Medium grey fields, with similarity indices between 67.55% and 79.86%, make up the next circle. Naturally, the values of indices descend as we progress eastwards. While C.27(*27*) and C.29(*29*) have 92.17% of their forms in common, C.29(*29*) has a 79.57% resemblance with C.39(*39*), 70% with C.36(*36*), 65.65% with L.24(*73*), 56.09% with Le.2(*161*) and finally reaches the aforementioned 41.74% with Le.5(*164*). What we do not obtain is an indication of the way in which the dialects resemble or differ from each other, given that the whole point of the method is to eliminate the qualitative component.

Fig. 2 exhibits the results of a comparison of the data for P.4(*123*), selected because it is the point with the highest similarity index, with the data for everywhere else. At first sight, this may appear to be a similar kind of map to the previous one, except that the point being compared with all others is further east. In the first choropleth map, for C.29(*29*), we observed a well-defined western area in A Coruña province. What we find now is a central area in the province spreading southwards into the north of Pontevedra province, with dark grey flanking this light grey and medium grey area on either side. However, we should note the indices in the legend:

---

[10] The visualization algorithm used for maps 1–4 is MINMWMAX. Visualization algorithms serve to "convertir la variation numérique (continue ou quasi-continue) en une variation iconique discrète" (Goebl 1987: 79), so they create groupings of indices by breaking the continuum of indices into intervals and assigning a colour to each interval. MINMWMAX "engendre un tissu polyogonal peu accidenté" (p. 89) because it distributes the indices smoothly along the scale; consequently, it "rehausse les macro-structures d'une distribution dialectométrique à visualizer" (p. 90). Given the high similarity scores of the data, I have decided to generate four-coloured choropleths, because four colours suffice to show the structure of the Galician dialects, and given that the choropleths must be generated in black and white, more than four groupings might be difficult to read.

[11] I will use four shades of grey for maps 1–8: light and medium grey (equivalent to the hot colours red and orange, respectively), and dark grey and very dark grey (equivalent to the cold colours green and blue).

	fig. 1: C.29(*29*)	fig. 2: P.4(*123*)
Light grey	79.87% — 92.17%	87.1% — 95.65%
Medium grey	67.55% — 79.86%	78.55% — 87.09%

The light grey colours in fig. 2 for P.4(*123*) only begin to be applied when similarity indices reach 87.1%, whereas in C.29(*29*) they are applied from 79.87% up. We can also compare the indices for the medium grey in fig. 1 for C.29(*29*) with the medium grey in fig. 2 for P.4(*123*). It can be clearly seen that P.4(*123*) has higher similarity indices with other localities, reaching an index of 95.65% in comparison with P.3(*122*), the highest such index in the entire database. The similarity index between P.4(*123*) and Le.5(*164*) is 51.74%, as compared to 41.74% for C.29(*29*).

This fact is obviously related to another: in fig. 1, P.4(*123*) shows up as medium grey vis-à-vis C.29(*29*), since their similarity level is 77.83%. Howwever, in fig. 2, C.29(*29*) shows as dark grey vis-à-vis P.4(*123*). The explanation is simply that in fig. 1, 77.83% is high enough for an index to be displayed as medium grey in relation to C.29(*29*), because the similarity indices obtained for C.29(*29*) are lower overall, but it is not high enough to make it into the medium grey in fig. 2 relative to P.4(*123*) because this choropleth has higher similarity indices all round. Given that P.4(*123*) has higher similarity indices with the other points, indices need to reach 78.55% for the colour-assigning algorithm to assign medium grey. The similarity indices in P.4(*123*) are relatively higher than those in C.29(*29*). The mean value of indices for P.4(*123*) is 78.7%, with a standard deviation of 7.21, whereas the mean value of indices for C.29 is 67.54%, with a standard deviation of 9.87. Here, then, we have lower similarity indices coupled with higher standard deviation. The pattern seen in P.4(*123*) is more common than that found in C.29(*29*). This accounts for the affirmation that fig. 1 is not quite comparable, strictly speaking, to fig. 2.

Now, compare maps 3 and 4, which present two localities at opposite ends of the Galician language area, selected precisely for their geographical position and their similarity indices vis-à-vis C.29(*29*) and P.4(*123*). Both show which points are most similar and most different, as we would expect. P.29(*148*) shows a western, somewhat southern-skewed profile, while A.7(*159*) shows one that is eastern, and somewhat northern-leaning. Notwithstanding this, as in the preceding maps, and coinciding with what the Galician qualitative dialectology developed by Fernández Rei (1990) had already established, the coloured bands tend to run from north to south. But these two localities have a number of significant elements in common:

	fig. 3 P.29(*148*)	fig. 4 A.7(*159*)	fig. 1 C.29(*29*)	fig. 2 P.4(*123*)
Light grey	75.99%–84.78%	74.96%–85.65%	79.87%–92.17%	87.1%–95.65%
Medium grey	67.20%–75.98%	64.25%–74.95%	67.559%–79.86%	78.55%–87.09%

As the table shows, the numbers in the maps for A.7(*159*) and P.29(*148*), which are quite similar, in general show lower similarity indices with other localities than C.29(*29*) and P.4(*123*), which appear to be more closely integrated into the Galician dialect map. This is shown by another comparison between fig. 2 and fig. 3. In fig. 2, locality C.27(*27*) is coloured medium grey with respect to P.4(*123*), because its similarity index is 79.13%. In fig. 3, C.27(*27*) is shown in light grey in relation to P.29(*148*), which might seem surprising since these points are relatively remote from each other. C.27(*27*) and P.29(*148*) have a similarity rate of 77.39%. As can be seen, the index of 79.13% in fig. 2 gets coloured medium grey, whereas a lower index of 77.39% gets coloured light grey in fig. 3.

Given that the points plotted do not have equal similarity indices, we might say that in *ALGa* there are localities that are more similar to each other and others that resemble each other to a lesser degree overall. Not all points are relatively the same or different, for if they were we might expect P.4(*123*), C.29(*29*), A.7(*159*) and P.29(*148*) to show similar maxima of similarity indices and a similar distribution of the indices that determine the colours. That this is not the case is demonstrated by the fact that in fig. 2, for P.4(*123*), the limits for maxima of similarity, represented by light grey, oscillate between 87.10% and 95.65%, whereas in fig. 3, for P.29(*148*), maximal indices range from 75.99% to 84.78%. In contrast to the index mean of 78.7% with a standard deviation of 7.21 for P.4(*123*), P.29(*148*) has a mean similarity index of 67.19% with a standard deviation of 7.54 and A.7(*159*) has a mean similarity index of 64.24% with a standard deviation of 7.43. In conclusion, some points are more similar to each other than others.

The different behaviour of different points is confirmed by a histogram analysis (I owe this suggestion to João Saramago):

1. At P.4(*123*), 106 points have above-mean similarity values.
2. At C.29(*29*), 76 points have above-mean similarity values; at P.29(*148*), 82; at A.7(*159*), 86.

P.4(*123*) not only has higher similarity values but of the 166 points with which we have compared it, 106 or 63.85% have an above-mean resemblance to P4(*123*); whereas C.29(*29*) has 45.78% above-mean values, P.29(*148*) has 49.39% and A.7(*159*) has 51.80%. Thus C.29(*29*), P.29(*148*) and A.7(*159*) are less highly integrated than P.4(*123*) in as much as they

have fewer similarities with other localities. In the next section, we shall see how this situation is accounted for.

6.2. *Maxima of similarity (see fig. 5)*

The maxima of similarity choropleth is obtained by selecting, for each locality in the database in turn, the highest similarity index found with another locality. Light and medium greys are assigned to places with higher values, and dark and very dark greys to those with lower values. Goebl (1993: 288) considers the light grey points, with high maxima of similarity indices, to mean dialect kernels, which are surrounded by other less homogeneous areas.

Fig. 5, obtained with an algorithm[12] which produces a four-colour figure reveals several dialect kernels with varying degrees of contiguity. The most striking grouping is encountered in the north of Lugo, starting at Cervo and spreading up the Northern Serras. This is the most cohesive area in the database. Each such point has similarity indices with some other point exceeding 93.83%, so for example L.5(*54*) is as high as 94.78% with L.2(*51*), L.7(*56*), L.9(*58*) and L.12(*61*). Also striking is the sharp drop to the east of the area, across the river Eo where eastern Galician commences; and to the south, in the direction of Castroverde, Lugo and Outeiro de Rei. The whole very dark grey area of eastern and Asturias Galician fits in with what we saw earlier in connection with fig. 4.

Near the Arousa estuary and along the river Ulla, we find another dialectal kernel bordered by territories of the centre located on the shore of the Deza and, particularly, south of the mouth of the Ulla. There seems to be a fair degree of affinity to the Ulla-Arousa kernel on the Muros estuary and the coast all the way to Fisterra, resulting in greater homogeneity to the northwest of the Ulla than to the east and south. Separate from, but very near this Ulla-Arousa kernel is another kernel centred around the Deza region in central Galicia. This fits the choropleths in maps 1, 2 and 3.

Another, less intense kernel shows up around Carballo. Dropping off gently, it is connected via Mariña (in A Coruña province) to the area in Lugo province mentioned above. It is also connected via the aforementioned Deza region kernel, along an elongated medium grey strip, to other homogeneous blocks located along the west bank of the Miño and south of

---

[12] Maps 5, 6, 7, 8, and 9 were created with the MEDMW visualization algorithm. This algorithm produces sharper profiles than MINMWMAX, "en accentuent les microstructures tout en fournissant des reliefs plus tourmentés" (Goebl 1987: 90). MEDMW usually gives a wider range of values corresponding to extreme intervals, distancing these from the central ones. A comparison between the choropleths obtained with MEDMW and MINMWMAX does not come within the scope of the present paper.

the Sil, north of Serra de Queixa and east of Serra do Courel. Less compact, the area is prolonged to the south of the river Sil through the centre of Ourense province until it reaches the source of the Limia in the west, and the country around Xinzo, Laza and Monterrei.

These are the main concentrations of similarity in the Galician linguistic territory insofar as phonetic data are concerned. We may discern a geographical centre displaying light and medium grey, which tends towards unity in the shape of these kernels, with a less uniform periphery. One might wonder whether such a configuration can be accounted for in terms of neolinguistic wave theory.

Le.4(*163*) and Le.5(*164*) are the least integrated areas, with lower indices than in the rest of the Galician language area. The whole eastern region from Lugo and Ourense on, including Asturias, León and Zamora, as well as the southwestern triangle of Pontevedra province, display lower rates of similarity. Now it is clearer why A.7(*159*) and P.29(*148*) should have lower similarity values with their neighbours than P.4(*123*) since, as the preceding discussion showed, these areas are less uniform than the rest of the Galician language area.

Before concluding this section, a glance at the border territories is in order. Much has been written about the Galician-Leonese border. It is well known that Menéndez Pidal (1906) proposed taking the isogloss for diphthongization of Latin stressed ŏ as the borderline dividing Galician-Portuguese from what he called Central Ibero-Romance, in which he included Leonese, Castilian and Aragonese. Of course such a division is conventional; what we really find in the north of the Iberian Peninsula is a geolectal continuum without clear-cut boundaries. As Aurrekoetxea (2010: 58) observes, in traditional dialectology some linguistic features are accorded more importance than others, perhaps as a result of an atomistic perspective according to which "a language is a set of features and characteristics that may be treated individually as if each functioned in isolation". Consequently, a hierarchy of features is posited for which "there is no specification of exactly what type of characteristic should be considered the most important or major one". But as Aurrekoetxea later points out (Aurrekoetxea 2010: 59; my translation from Spanish):

> Focusing on systematic features only, each of these performs a particular function in the system that only it can perform, such that, if the feature were to change or disappear, restructuring of the system would ensue. It follows that all features of a language system are equally important.

One aspect of fig. 5 that has already been mentioned is the high degree of similarity between the *ALGa* locations in phonetic data: even the lowest index found is 71.3% for data shared by Le.5(*164*) and Le.4(*163*). This is

really high, bearing in mind that traditional dialectology does not even classify Le.5(*164*) as Galician, but Leonese, since it conserves Latin intervocalic N and diphthongizes Ŏ. What is more, Le.5(*164*) has the following similarity indices in comparison with some places traditionally considered Galician-speaking:

63.04% with A.6(*158*) and A.7(*159*)
62.61% with L.23(*72*) and O.4(*92*)

It actually reaches indices of 60% or more in 19 places. These indices of Le.5(*164*) stand in contrast to those obtained with places in western A Coruña province:

41.74% with C.29(*29*)
42.17% with C.33(*33*)
43.28% with C.38(*38*) and C.27(*27*)

It is instructive to compare the indices for Le.4(*163*), which while traditionally considered Galician is quantitatively the closest to Leonese Le.5(*164*), with the same points in western *Coruñés*:

C.29(*29*): 51.74%
C.33(*33*): 53.91%
C.12(*12*): 54.35%
C.27(*27*): 54.70%

As we can see, the differences are always in the same direction but at a separation of 10 percentage points. However, three facts that in my view are of fundamental importance should be considered here:

a) Although the point closest to Le.5(*164*) is Le.4(*163*), the point closest to the latter is O.3(*91*); indeed, 39 localities are closer to Le.4(*163*) than Le.5(*164*) is.

b) Four other Galician-speaking localities share the same similarity index as Le.4(*163*) and Le.5(*164*), namely L.21(*70*), L.22(*71*), L.25(*75*) and O.14(*102*).

c) There are no fewer than 123 Galician-speaking localities with similarity indices with Le.4(*163*) lower than the 71.3% found between Le.4(*163*) and Le.5(*164*).

Thus, the Galician-speaking locality Le.4(*163*) presents 123 Galician-speaking areas with which its similarity indices are lower than the similarity indices it shares with a Leonese-speaking locality! What is more: a Galician locality such as L.21(*70*) has a similarity index of 55.22% with the Leonese locality Le.5(*164*), for example, as compared to similarity indices of 57.83% with C.29(*29*), or 57.39% with P.29(*148*).

For one thing, this is a nice illustration of a geolectal continuum. It is also a sobering example for anyone who believes that a single linguistic

phenomenon can be chosen to mark the frontier between two historically established languages. Two dialects of a single historical language, such as Le.4(*163*) and C.29(*29*), may have more differences than a dialect of one historical language and a dialect of a different historical language, such as Le.4(*163*) and Le.5(*164*). It should be noted that the similarities calculated for all these localities do not depend on just two features, diphthongization of Ŏ and loss of intervocalic N, but on the features covered by 230 working maps, none of which is hierarchically classified as more or less important than any other.

6.3. *Means of similarity and standard deviation (see figs. 6 and 7)*

Fig. 6 classifies localities in terms of each one's mean similarity index. This map promises to be of interest because "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).

Fig. 7 shows the standard deviation for each mean similarity. It is useful to look at both maps together so that we can observe both the mean similarity index and the standard deviation for each locality. This information is mutually complementary, because a mean may either reflect a compact concentration of values that are close to each other (i.e. a low standard deviation, with greater similarity within the sample) or conceal dispersed values that differ substantially from the mean (i.e. a high standard deviation, with big differences within the sample).

As fig. 6 shows, the whole of central Galicia has relatively high mean similarity indices, whereas the periphery has lower values. The high mean values signify high similarity indices (perhaps with low standard deviation). But the low mean values may result either from generally lower similarity indices (with low standard deviation) or from heterogeneous groups and both high and low values (with high standard deviation).

The location of the high means, shown in light grey, corroborates the pattern of greater homogeneity in central Galicia with its high mean similarity and low standard deviation, as indicated by the very dark grey and dark grey colours in fig. 7. This tells us that most of the similarity indices for any given locality have values close to the mean value for that locality, and since these locations all have high mean values and low standard deviation, the conclusion is that the localities in this area have many features in common. For the very reason that they have so many features in common, such areas of high mean indices and low standard deviation are

zones of transition[13] between different language varieties with more marked characteristics.

By way of contrast, the west of A Coruña province presents relatively low mean similarity indices (coloured very dark grey in fig. 6) and very high indices of standard deviation (in light grey in fig. 7): just the opposite of what we just saw in central Galicia. These places are less similar to other localities (as is also evident, in fig. 5), given the low mean similarity values; moreover, standard deviation is high because some similarity indices in these localities differ more from their mean, sharing low values with some places (from which they are therefore more different) and higher ones with others (to which they are more similar). Northeastern Galician presents a comparable but more marked pattern, subject to a similar interpretation: some localities are very similar to each other but highly distinct from the remainder.

The southwest of Pontevedra province and the Galician spoken in Zamora and southern León are interesting. Mean similarity is fairly low (appearing in very dark grey in fig. 6), but so are the standard deviation indices (very dark grey in fig. 7). That means that the mean similarity levels are reliable in the sense that each locality consistently has relatively low similarity indices with everywhere else. If that is the case, then these dialects are even more distinctive and independent than Asturian and northwestern varieties, having low similarity indices with every other place covered by *ALGa.*

To recapitulate: the comparison of mean similarity with standard deviations yields three identifiable patterns: high means with low standard deviation in central Galicia, low means with high standard deviation in the northwest and northeast, and, lastly, low means with low standard deviation in the southwest and southeast.

6.4. *Maps of the asymmetry coefficient of similarity distribution (see fig. 8)*

Finally, fig. 8 shows an asymmetry coefficient (skewness) choropleth. To interpret it, first let us look at the histogram in fig. 1. The mean of all similarity values with C.29(*29*) is 67.54%: 90 places have similarity indices with C29(*29*) below the mean, while 76 have similarity values above the mean. Thus, there are more localities that are below the mean (90) than above it (76). We can say that this is a left-leaning asymmetrical distribution, skewing to the left. However, in fig. 2 we have the opposite distri-

[13] "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).

bution. With P4(*123*), the mean of similarity is 78.54%: 70 points have indices of similarity below the mean and 96 above. In this case, we can say that it is a right-leaning asymmetrical distribution, skewing to the right.

According to Goebl (1981: 401; 1992: 448–449; 2007: 141–142), a right-leaning distribution facilitates communication among speakers because it means that, for the most part, speakers share the same features and the differences are minor for most places. This makes sense given that most places are above the mean (i.e. there are more localities with more similarity between them than places with less similarity).

To measure the asymmetry, skewness, of a distribution, Goebl (1981: 400–406; 1992: 449; 1993: 82) uses a scale called *Fisher's asymmetry coefficient*. This assigns a value of 0 to symmetrical distributions (i.e. those with the same number of items above and below the mean), *positive* values to left-leaning, skewing to the left, asymmetrical distributions (where more localities have values below the mean, as in the histogram in fig. 1), and *negative* values to right-leaning asymmetrical ones (with more places having values above the mean, as in the histogram in fig. 2). In the choropleth of fig. 8, very dark grey is assigned to localities with negative values and light grey to places with positive values.[14] Goebl (1992: 449) says:

> Aux valeurs positives du coefficient de Fisher correspondent surtout les dialecticités mal insérées dans l'ensemble du réseau étudié. *Sociologiquement* parlant, il s'agit de personnes peu capables ou guère désireuses d'interagir d'une façon positive avec le reste du groupe auquel elles appartiennent ("trouble-fête", "boucs émissaires"). Les dialecticités symbolisées en bleu foncé (etc.), par contre, correspondent à des individus dont les dispositions interactives sont grandes ("médiateurs", "boute-en-train", etc.). D'un point de vue dialectologique, l'on peut associer la notion de "zone conservatrice" ou de "résidu" aux espaces rouges [...], et la notion de "zone de transition" ou d'"amphizone" aux secteurs bleus.

Thus light grey areas in choropleth 8, which share fewer features with the rest because they are higher than the mean (with positive asymmetry values) are those which are conservative and less closely linked to other localities, while very dark grey areas (with negative asymmetry values in their similarity distribution, being higher than the mean) are places that are more innovative, integrated and receptive, with less marked peculiarities, given that their features are those which are widespread elsewhere in the

[14] It may seem counter-intuitive to assign *negative* values to the asymmetry coefficient for places above the mean and *positive* ones to those below it, yet this is how this coefficient works.

territory. This fits everything we have seen so far: we have a more uniform linguistic heartland where language varieties share and exchange more features, and a periphery (north-western and north-eastern Galician) that is less uniform, with a more marked personality and greater independence. As the tenets of traditional dialectology would lead us to expect, the most marked dialects are found in peripheral areas.

This view contrasts, at least in the domain of phonetics, with the old belief expressed since Santamarina (1982) that Galician is more conservative as we move further east and more innovative towards the west. This assumption has always seemed questionable to me, but it is necessary to note that in dialectometry *conservative* does not mean "more like Latin" or "having changed less since Latin" but rather "less prone to communicate and exchange features, standing apart".[15] This view of dialectometry is not strictly incompatible with Santamarina's historical perspective, and it would be interesting to compare on a map the indices of Castilianization to see whether phonetic conservatism correlates with lexical conservatism.

6.5. *Isoglosses (see fig. 9)*

Two points need to be made about the isogloss maps generated by the *VDM* programme. They represent neither isoglosses for individual linguistic features nor borders between languages or language varieties, but mark the degree of difference observed between *adjacent localities* in the network. Secondly, the isogloss maps are calculated not from similarity indices but from distance indices.

To generate the isogloss fig. 9, *VDM* calculated 430 distance indices, this being the total number of borders between localities appearing on the map. In this instance, lighter grey colours stand for lower distance indices and darker grey colours for greater distances. The thickness of the line is another way to mark distance: thin lines for lower distance indices and thick lines for higher ones. This figure complements the previous ones and confirms the same pattern already indicated, namely that there is greater uniformity in the central areas and less on the periphery. The dialect kernels in the north of Lugo, the Carballo region, the mouth of the Ulla, and south of the confluence of the rivers Miño and Sil are still there, with low distance indices.

---

[15] Areas in light and medium grey "ont de mauvaises relations linguistiques avec les parlers voisins et conservent bon nombre de caractéristiques propres. Inversement, en jaune et bleu, les parlers [...] dont de bonnes relations avec les autres parlers et en même temps peu de caractéristiques propres". (Aurrekoetxea & Videgain 2009: 103)

However, a closer look at the histogram will show that the borders between localities with low distance indices (those between places with more similarity) number 302 out of the 430; that is to say, 70.23% of the borders have below-mean distance indices. In contrast, 128 borders present above-mean differences (indicating more sharply differentiated adjacent localities); so only 29.77% of the borders have high distance indices.

Although *high* distance indices (above the mean, shown in darker grey) are ones ranging from 15.22% to 18.25%, these cannot be said to be truly high indices. In fact, the whole distance index range represented by the five first divisions is from 4.35% to 18.26%, a range of just 13.91 percentage points. Thus, 380 of 430 borders present distance indices below 18.26%. Again, the analysis of the isogloss map serves to bear out the uniformity of the *ALGa*'s phonetic data.

6.6. *Grouping maps and dendrograms (see fig. 10a and b)*

Fig. 10 contains a choropleth map of groupings and a dendrogram. *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. The choropleth in 10 shows what happens when the programme is told to create five homogeneous dialect groups. The five groupings (called *choremes*)[16] are visible at a glance, but they do not comprise a flat structure: they do not all belong to the same level, nor are they equally interrelated, as the branches (*dendremes*) of the complementary dendrogram shows. As a matter of fact, underlying them is a sort of tree structure (hence the name dendrogram) in which elements only *appear* to be contiguous. Consider this analogy with syntax: adjacent words in discourse may belong to different syntactic units, so in *the man never sleeps*, the adverb *never* appears next to the noun *man*, yet the two words are only really associated via the intermediate structures *the man* on the one hand and *never sleeps* on the other.

The dendrogram informs us of the hierarchical structure of such groups or choremes in the choropleth. The localities in the darkest grey are directly

[16] The reasons for generating five groupings are the same as those for having four in the other maps: that they seem to suffice to show the dialectal structure of such a small territory with such high indices of similarity, and the fact that the choropleths must be in black and white. But there is a further reason which is linked to a point of curiosity. Given the question of the independence of Galician and Asturian Galician, it would be nice to know at what point the Asturian dialects of Galician broke off. Five groupings are needed to show the answer: Asturian Galician is more closely related to Central Galician than Central Galician is to Western Galician.

grouped together in opposition to the rest of the territory. Those coloured lighter grey stand against the medium grey, which form another block opposed to the lighter greys, while the darker greys are opposed to all these as a whole.

What this means is that the first major dialect division is that between western Galician and central-eastern Galician. The western dialect is divided in turn into two subdialects, an Atlantic coastal subdialect and an inland one. The central-eastern dialect is subdivided into a northeastern Galician subdialect and a central Galician subdialect, while the latter is further broken down into northern and southern varieties. Note the following:

1. All the lighter areas in northern Lugo province on the choropleth in fig. 5 belong to a single dialect.
2. Differentiation from this Lugo kernel is sharper to the east (with the river Eo constituting the boundary) than toward the south, implying an abrupt change of dialect.
3. The area of the Sil-Miño confluence is compact, in opposition to northern Lugo.
4. The Carballo area is linked to the Ulla area, but the latter does not continue along the Muros estuary, though it breaks abruptly in the centre with the Deza district and more gently with the whole south.
5. The entire transition area in the standard deviation fig. 7 is split between places that belong to the eastern part of the western macro-area and the west of the central-eastern macro-area.
6. Notwithstanding its "Leonese-ness", Le.5(*164*) still groups with central Galician, not with the north-eastern variety.
7. Asturian Galician is closer to Central Galician than Central Galician is to Western Galician. Thus Central Galician and Asturian Galician constitute a bloc opposed to Western Galician.

7. Conclusions

In the first part of this article we examined the database that serves as the basis for the creation of the maps. This was important in order for the reader to appreciate the significance of the data under review. It was explained that in the context of *ALGa* "phonetic data" should be understood as meaning the outcome of regular historical sound changes, involving differences that today might better be considered phonological, or more exactly phono-lexical. Attention was also paid to the database's actual structure, looking at the problems caused by gaps in the data, Spanish interference, divergences

in patterns of development, and the number and representativeness of the selected phenomena.

Next, keeping in mind all these considerations and caveats, the findings from the database were analysed through ten maps illustrating various perspectives on Galician phonetics, and it was shown that:

a) not all localities are equally distinct or different: some areas are more compact and show greater similarity, while others are less compact and display a greater degree of divergence;
b) there is a large, more uniform central area, with less internal variation, and there are more conservative peripheral areas with more sharply defined personalities and a lower degree of integration into the system as a whole;
c) by means of a standard deviation analysis, three types of area may be distinguished: a uniform central area, another at either end of the south, and a third one located in western A Coruña province and in the north-eastern corner of the Galician language area. Let it be noted that the aforementioned areas are not related to each other in terms of shared features but by the same patterning of linguistic facts.

Finally, the grouping choropleth and dendrogram demonstrate the internal structure of Galician dialects in phonetic terms, with a large western macro-dialect opposed to a large central-eastern macrodialect.

Bibliography

ALGa I = Instituto da Lingua Galega. 1990. *Atlas lingüístico Galego.* Vol. I. *Morfoloxía Verbal*. A Coruña: Fundación Barrié de la Maza.

ALGa II = Instituto da Lingua Galega. 1995. *Atlas lingüístico Galego.* Vol. II. *Morfoloxía Non-Verbal*. A Coruña: Fundación Barrié de la Maza.

ALGa III = Instituto da Lingua Galega. 1999. *Atlas lingüístico Galego*. Vol. III. *Fonética.* A Coruña: Fundación Barrié de la Maza.

ALGa V = Instituto da Lingua Galega. 2005. *Atlas lingüístico Galego.* Vol. V. *Léxico. O ser humano (I). Ser físico*. *Morfoloxía Verbal*. A Coruña: Fundación Barrié de la Maza.

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

Aurrekoetxea, Gotzon. 2010. Sobre la dialecticidad de los dialectos. In *Homenaxe al Professor Xosé Lluís García Arias*, Tomu I, 53–77. Uviéu: Academia de la Llingua Asturiana.

Aurrekoetxea, Gotzon & Charles Videgain. 2009. Le project Bourciez: traitement géolinguistique d'un corpus dialectal de 1895. *Dialectologia* 2, 81–111.

Fernández Rei, Francisco. 1990. *Dialectoloxía da lingua galega.* Vigo: Xerais.

Goebl, Hans. 1981. Éléments d'analyse dialectométrique (avec application à l'AIS). *Revue de Linguistique Romane* 45, 349–420.
Goebl, Hans. 1987. Poins chauds de l'analyse dialectométrique: pondération et visualisation. *Revue de Linguistique Romane* 51, 63–118.
Goebl, Hans. 1992. Problèmes et méthodes de la dialectometrie actuelle (avec application á l'AIS). *IKET-7, Actas del Congreso Internacional de Dialectología* (1991), 429–475. Bilbao: Euskaltzaindia.
Goebl, Hans. 1993. Dialectometry: A short overview of the principles and practice of quantitative classification of Linguistic Atlas Data. In Reinhard Köhler & Burghard B. Rieger (eds.), *Contributions to Quantitative Linguistics*, 277–315. Dordrecht: Kluwer.
Goebl, Hans. 2002. Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de linguistique romane* 66, 5–63.
Goebl, Hans. 2003. Regards dialectométriques sur les données de l'Atlas linguistique de la France (ALF): relations quantitatives et structures de profondeur. *Estudis Romànics* 25, 59–121.
Goebl, Hans. 2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21, 411–435.
Goebl, Hans. 2007. A bunch of dialectometric flowers: a brief introduction to dialectometry. In Ute Smit, Stefan Dollinger, Julia Hüttner, Guntger Kaltenböck, Ursula Lutzky (eds.), *Tracing English through time. Explorations in language variation.* Wien: Wilhem Braumüller Universitäts-Verlagsbuchhandlung.
Goebl, Hans. 2008. La dialettometrizzazione integrale dell'AIS. Presentazione dei primi risultati. *Revue de Linguistique Romane* 72, 25–113.
Labov, William. 1972. *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
Labraña Barrero, Sabela. 2009. Las consonantes fricativas de la lengua gallega. *Estudios de Fonética Experimental* 18, 193–213.
Martínez-Celdrán, Eugenio & Xosé Luís Regueira. 2008. Spirant approximants in Galician. *Journal of the International Phonetic Association* 38, 51–68.
Menéndez Pidal, Ramón. 1906. El dialecto leonés. *Revista de archivos*, *bibliotecas y museos* 1–2, 128–311.
Mohanan, K. P. 1995. The organization of the grammar. In J. Goldsmith (ed.) *The handbook of phonological theory*, 24–69. Oxford: Blackwell.
Rivadeneira, Marcela & Xavier Casassas. 2009. New insights into the use of VDM: some preliminary stages and a revisited case of dialectometry. *Dialectologia* 2, 23–35.
Santamarina, Antón. 1980. Dialectoloxía galega: historia e resultados. In Ramón Lorenzo & Dieter Kremer (eds.) *Tradición*, *actualidade e futuro do galego. Actas do coloquio de Tréveris*, 153–187. Santiago de Compostela: Xunta de Galicia.
Trubetzkoy, Nikolay. 1939. *Grundzüge der Phonologie* (Travaux du Cercle Linguistique de Prague 7). – English translation (quoted here): *Principles of Phonology.* Berkeley/Los Angeles: University of California Press, 1969.
Vidal Figueiroa, Tiago. 1993. Proposta descritiva das consoantes fricativas alveolodentais dos dialectos galegos. *Cadernos de lingua* 7, 5–26.

Francisco Dubert García • Facultade de Filoloxía • Avda. Castelao s/n •
15782 Santiago de Compostela, SPAIN • francisco.dubert@usc.es

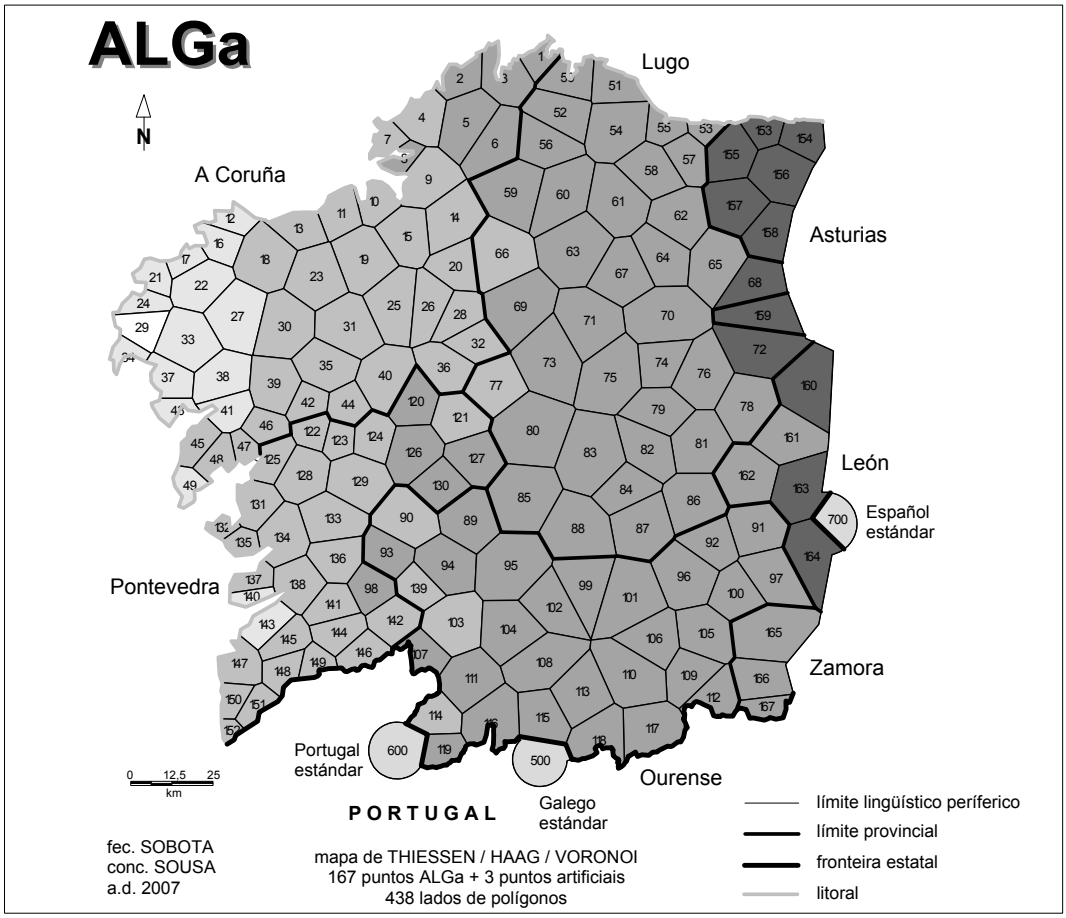
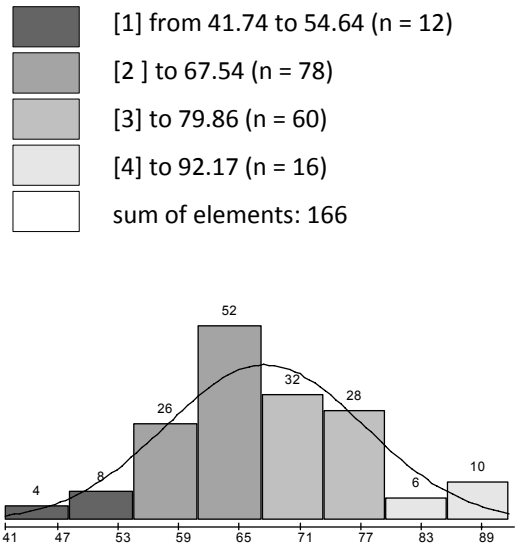


Fig. 1. C.29(*29*) compared to other *ALGa* localities



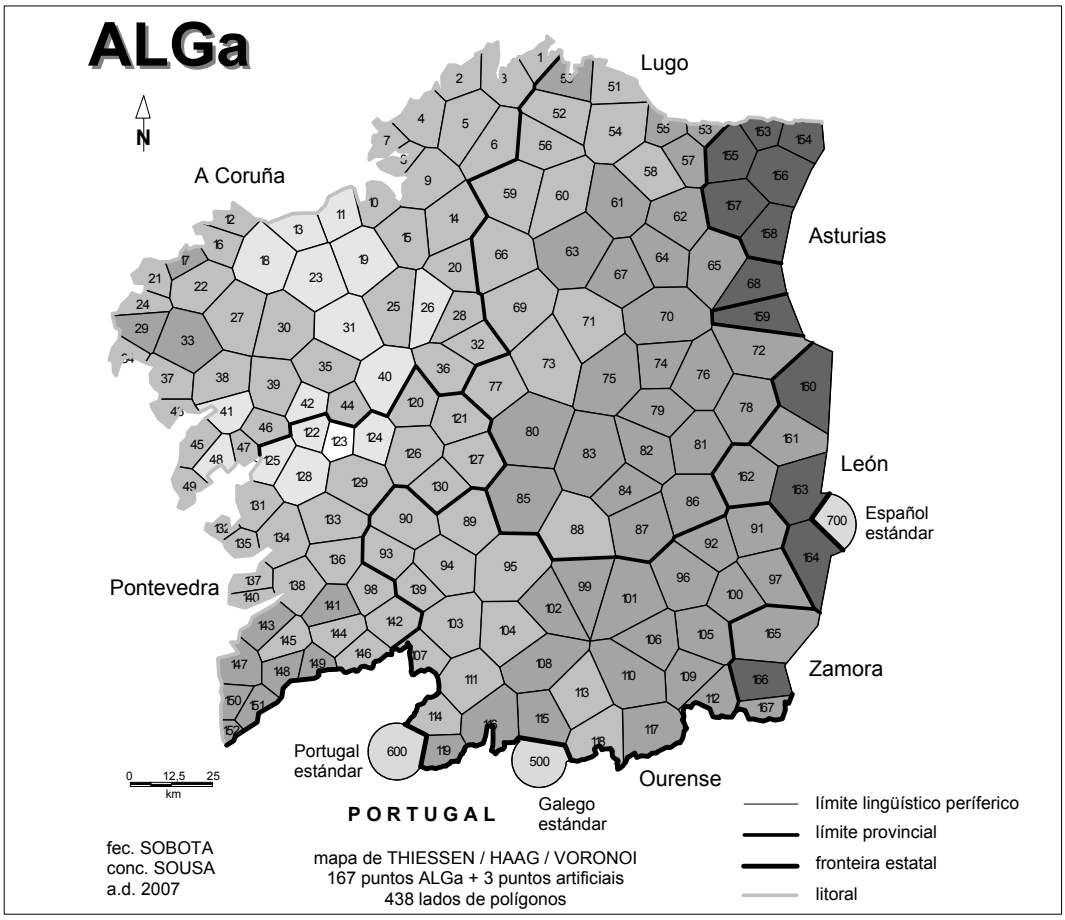
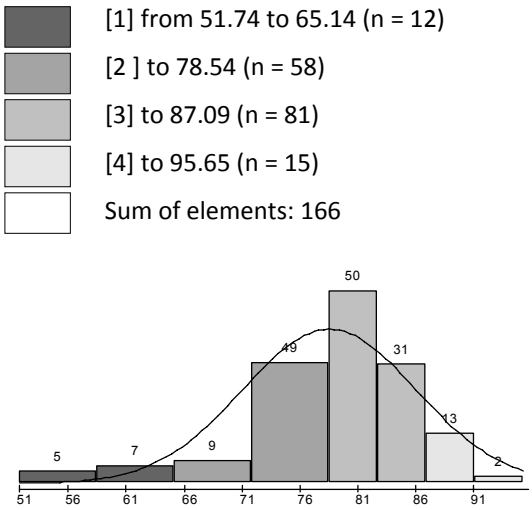


Fig. 2. P.4(*123*) compared to other *ALGa* localities



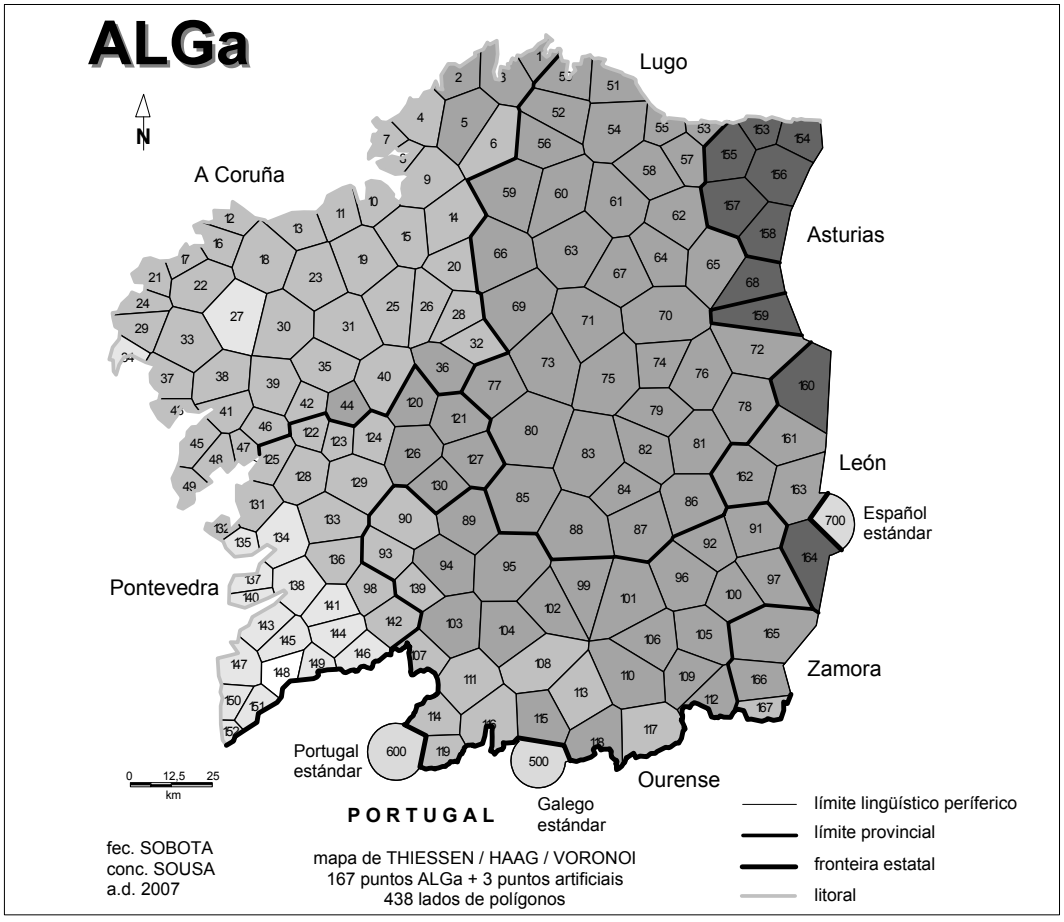
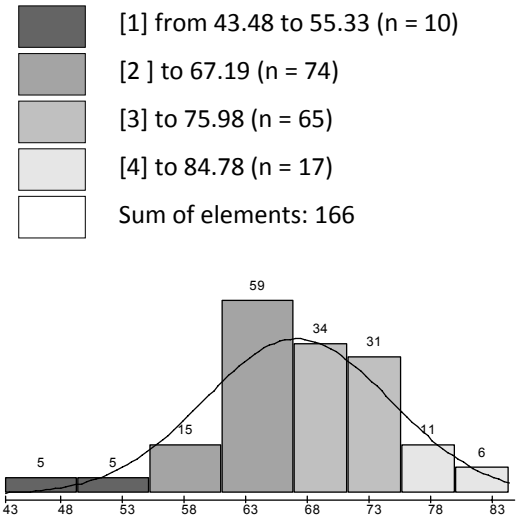


Fig. 3. P.29(*148*) compared to other *ALGa* localities



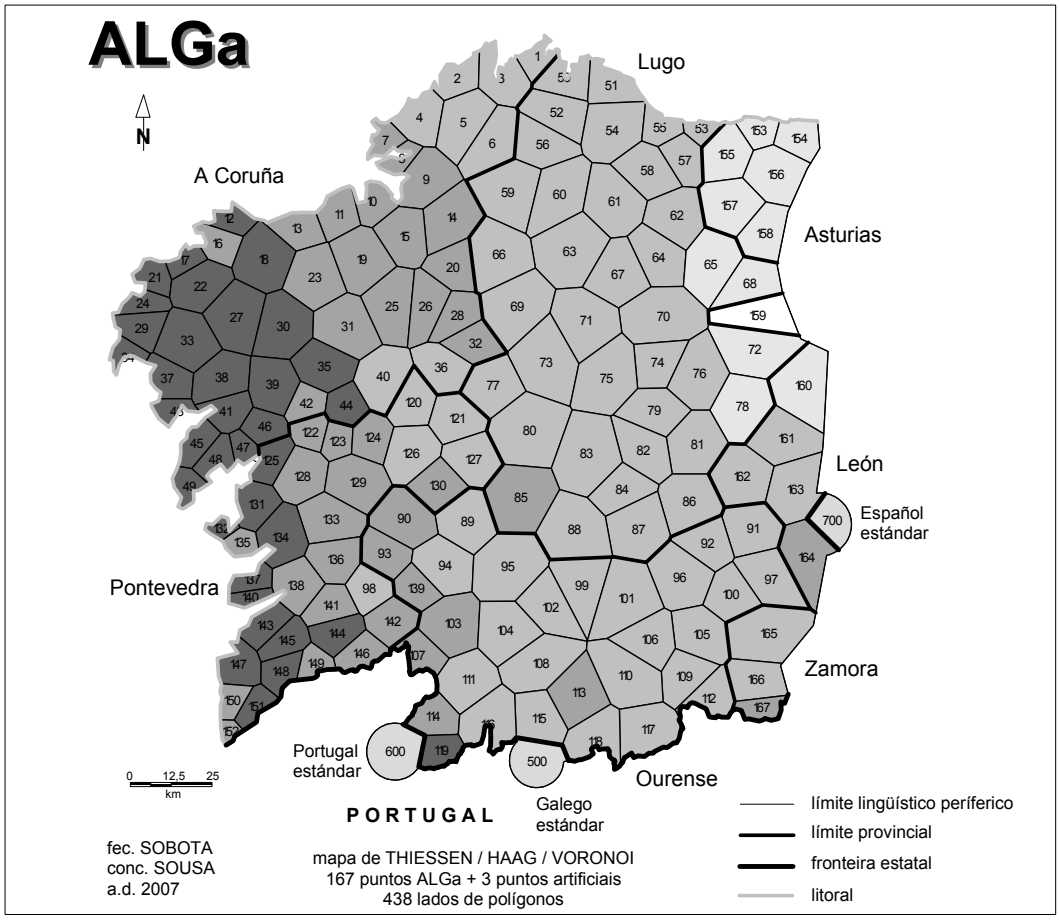
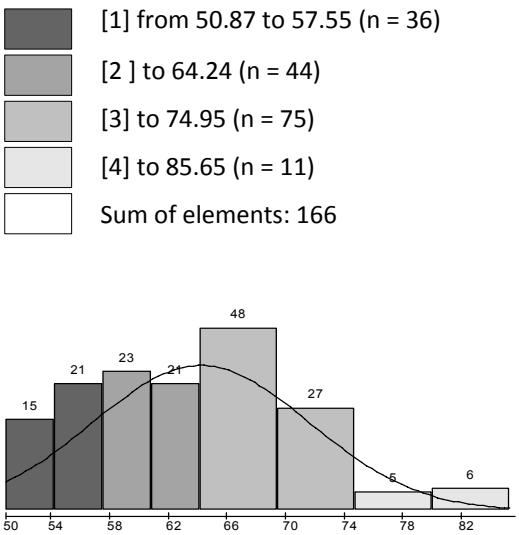


Fig. 4. A.7(*159*) compared to other *ALGa* localities



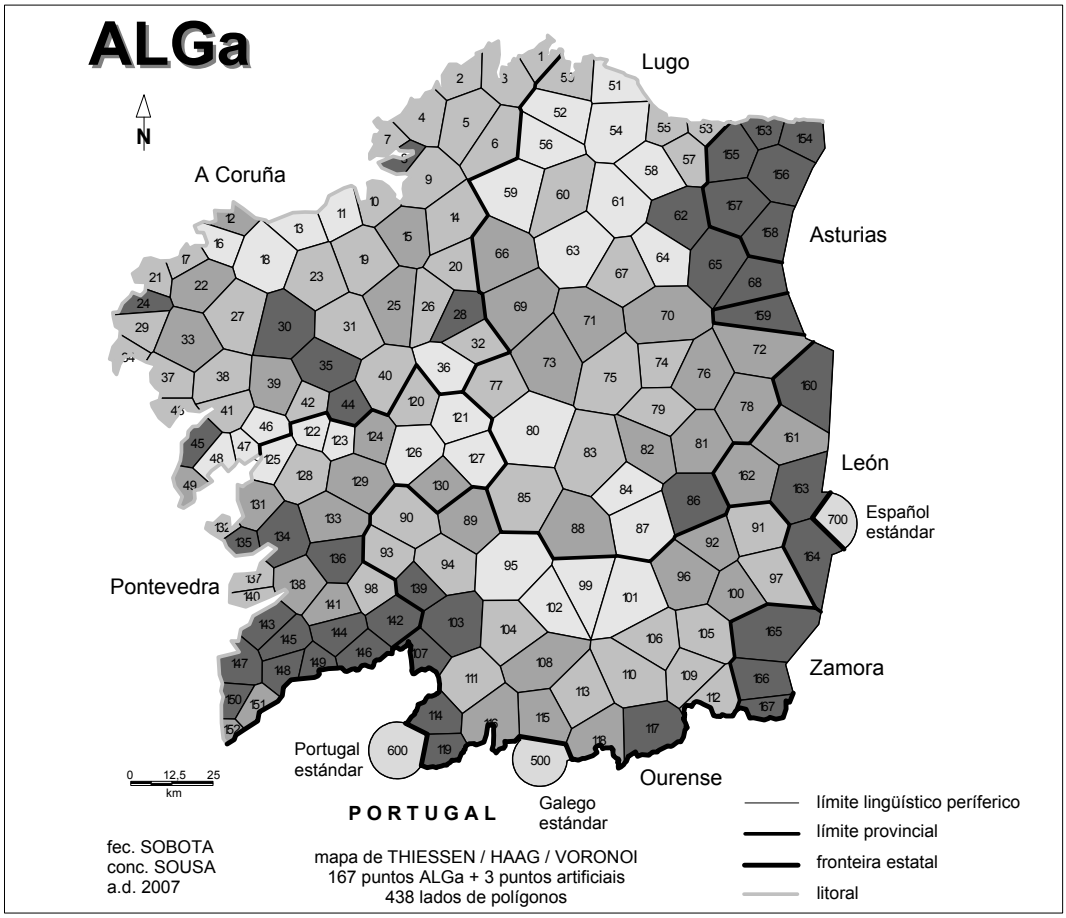
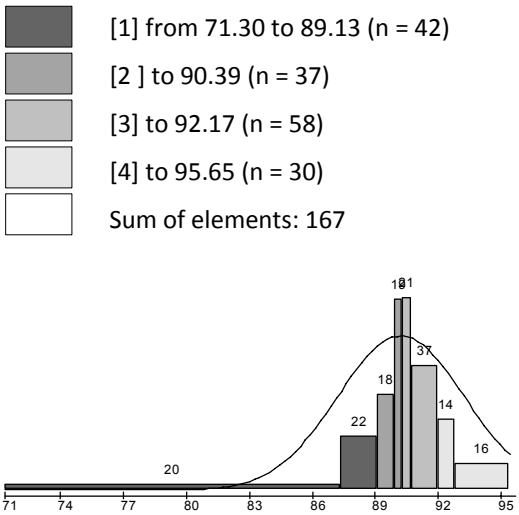


Fig. 5. Choropleth of maximum similarity indices using data from *ALGa*



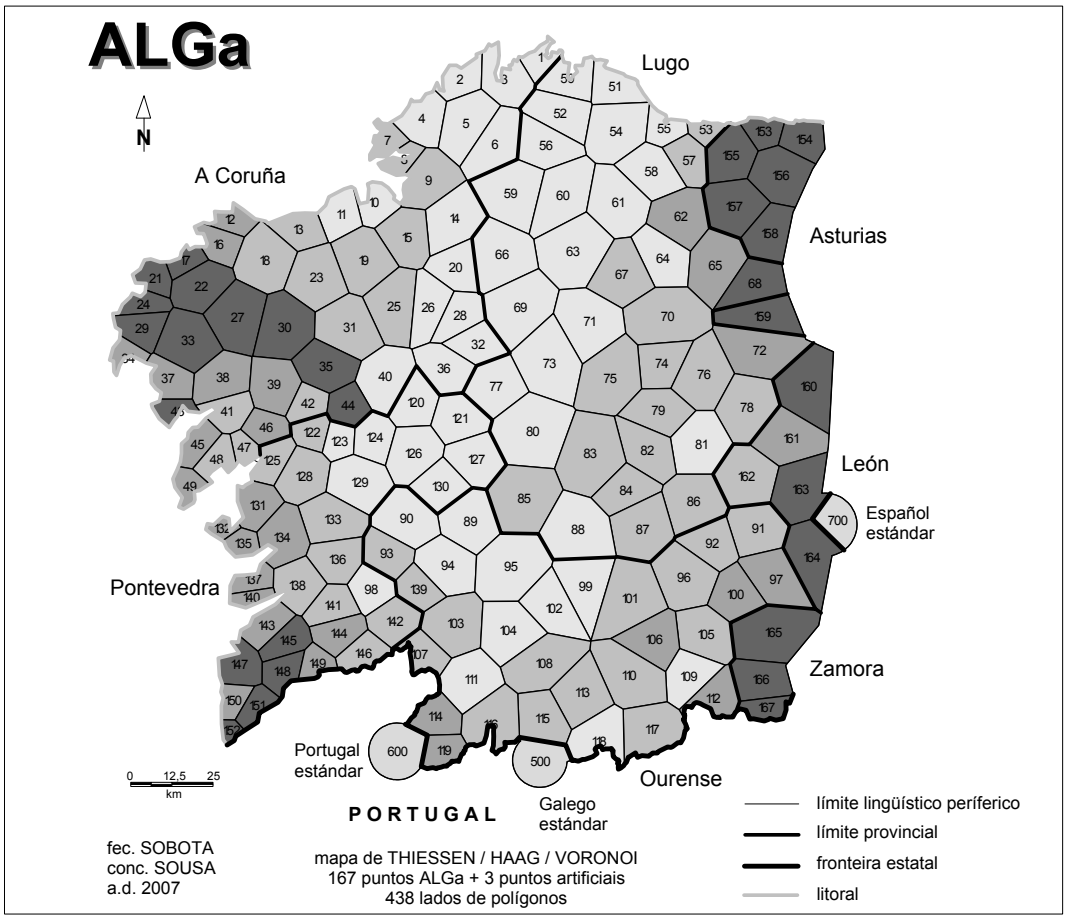
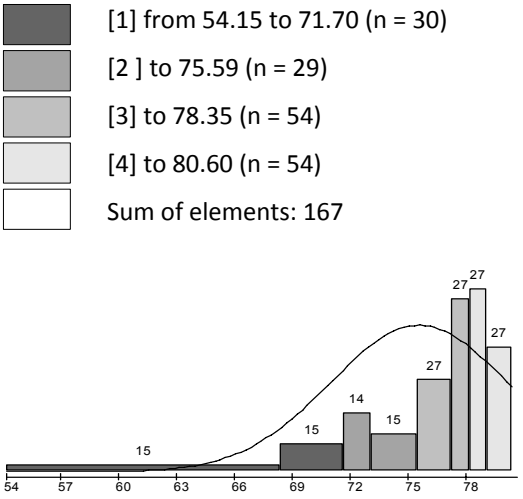


Fig. 6. Mean similarity indices for each locality



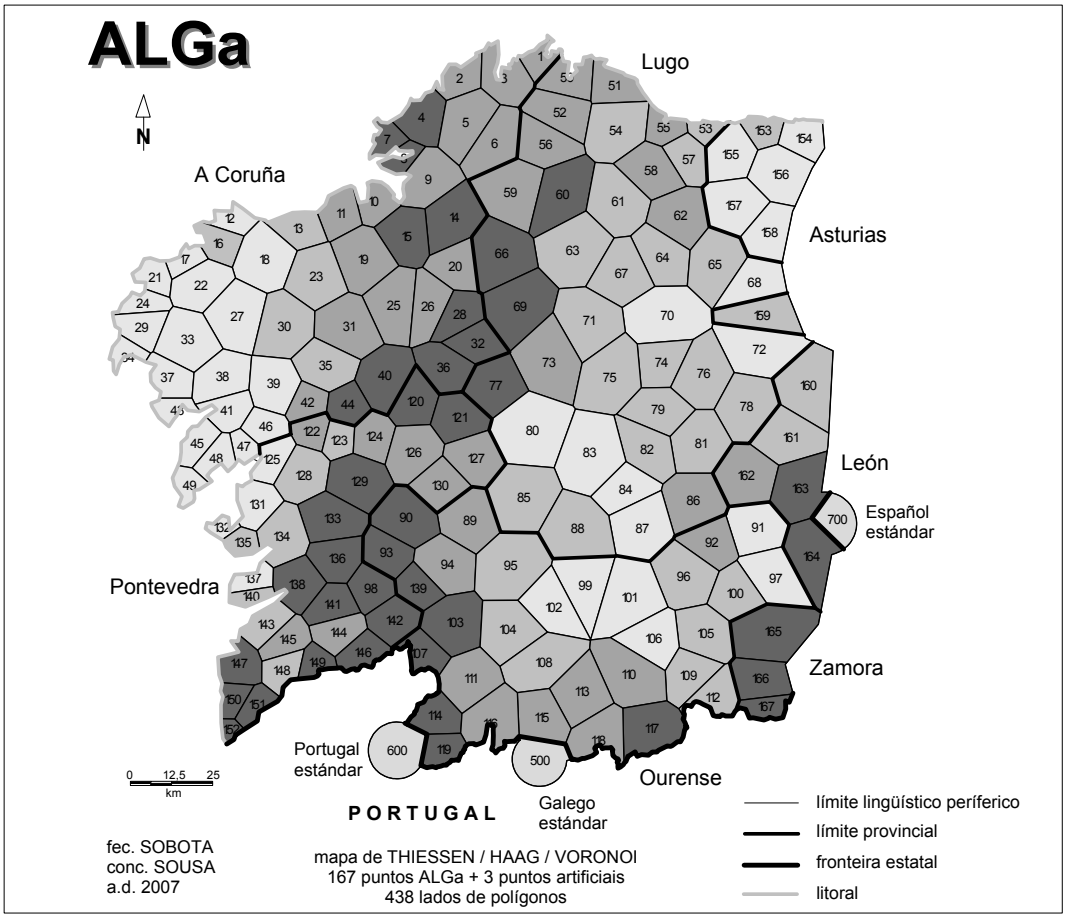
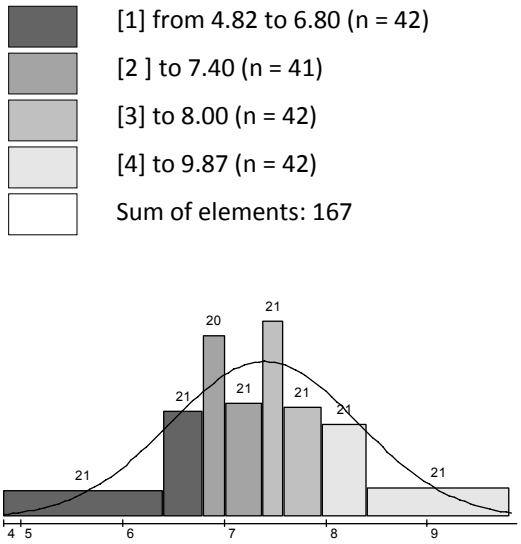


Fig. 7. Standard deviation indices



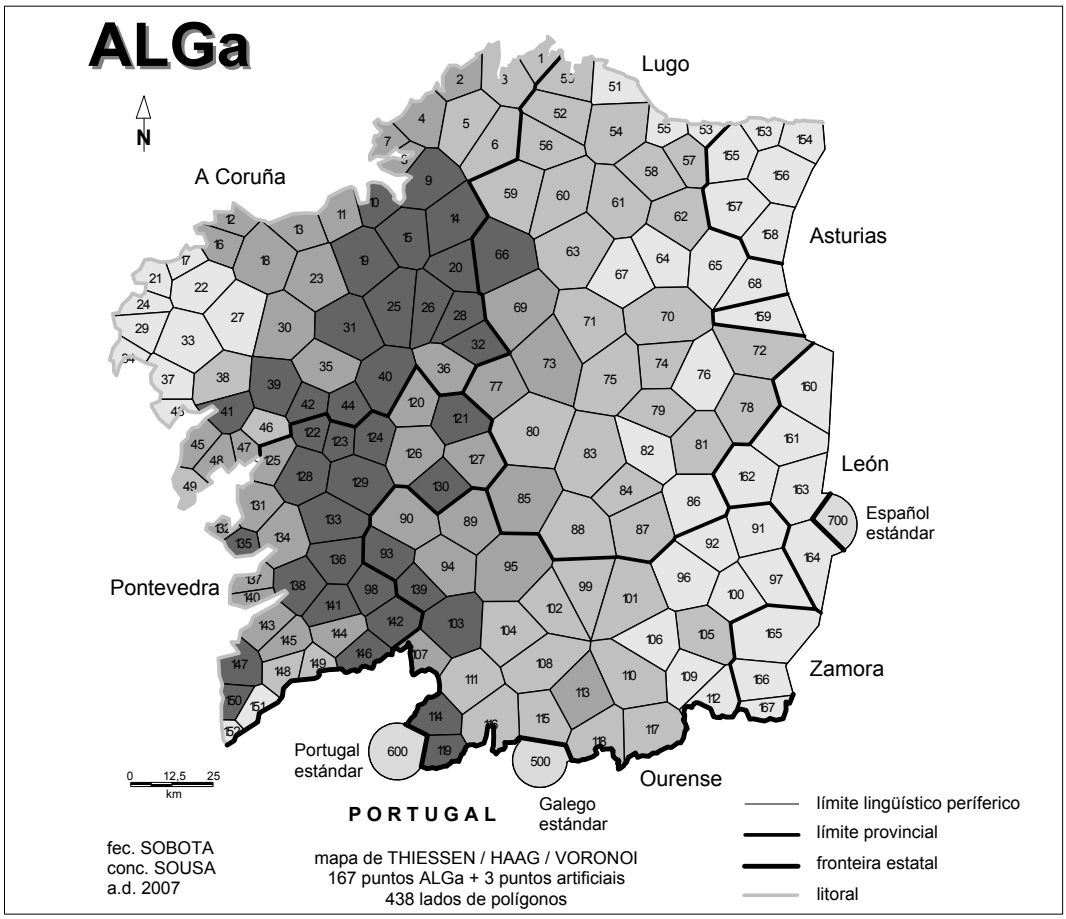
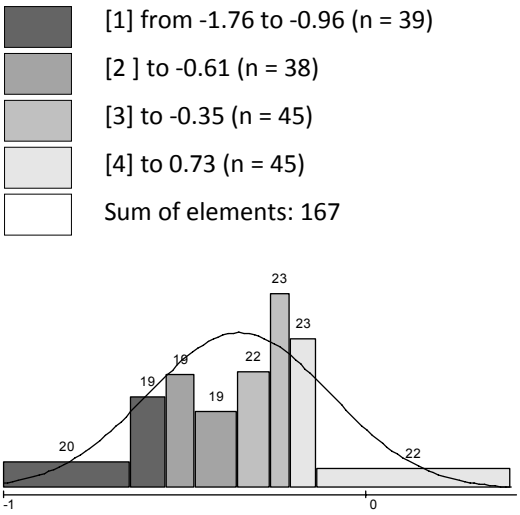


Fig. 8. Synopsis of asymmetry coefficient



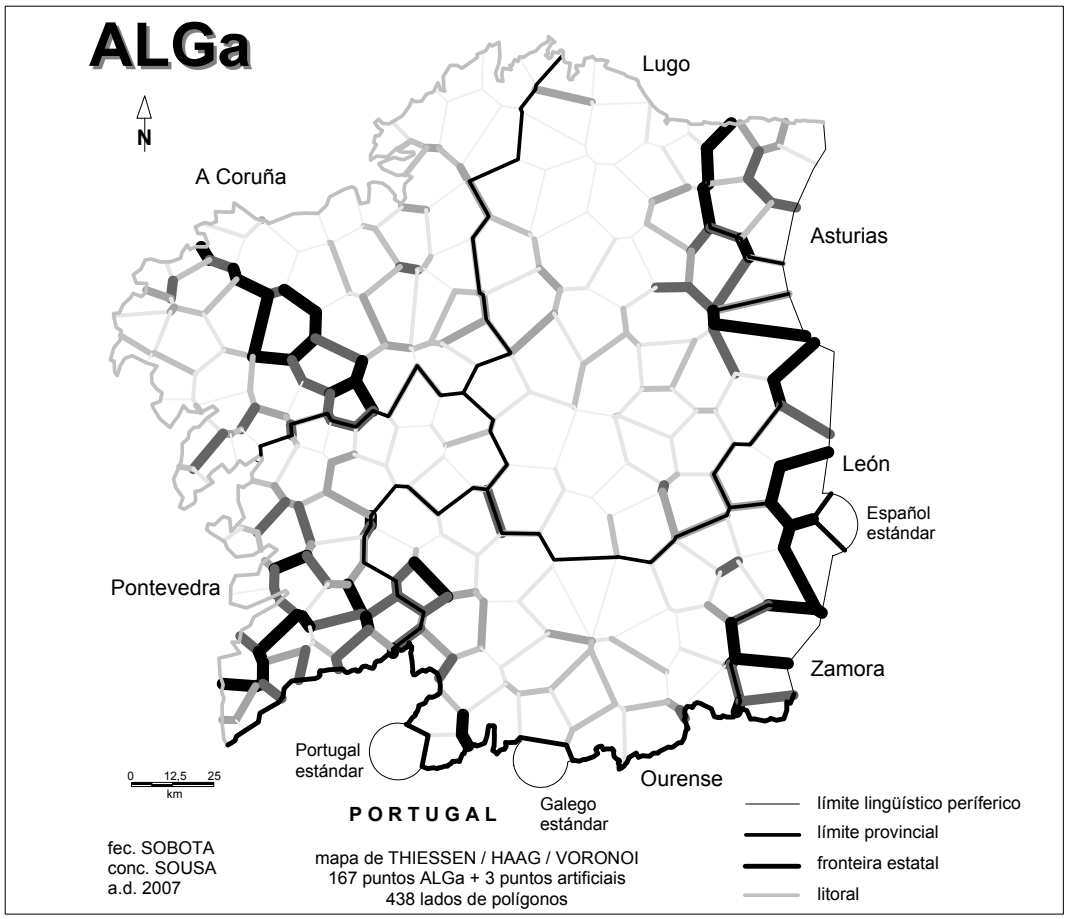
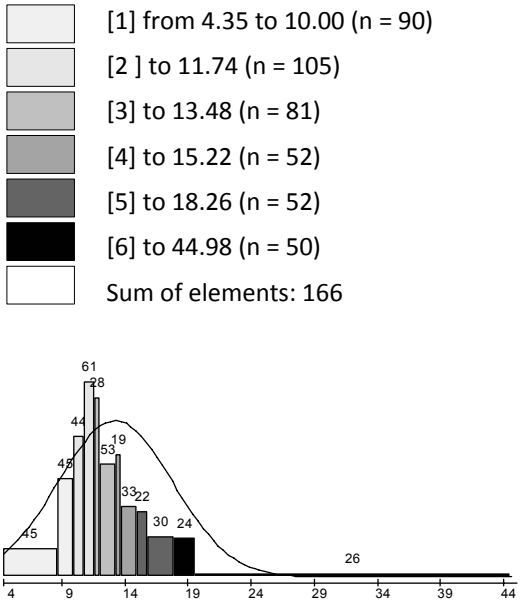


Fig. 9. Phonetic isoglosses for localities in *ALGa*



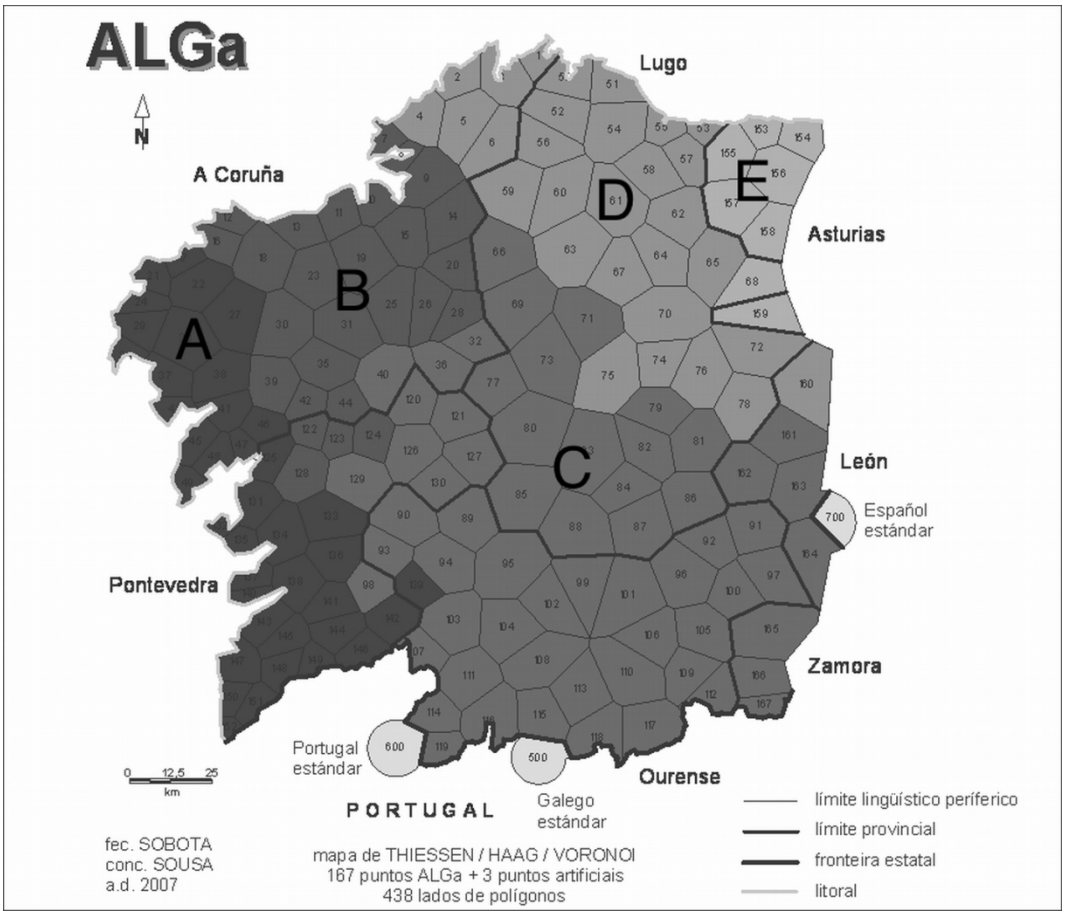


Fig. 10a. Grouping map

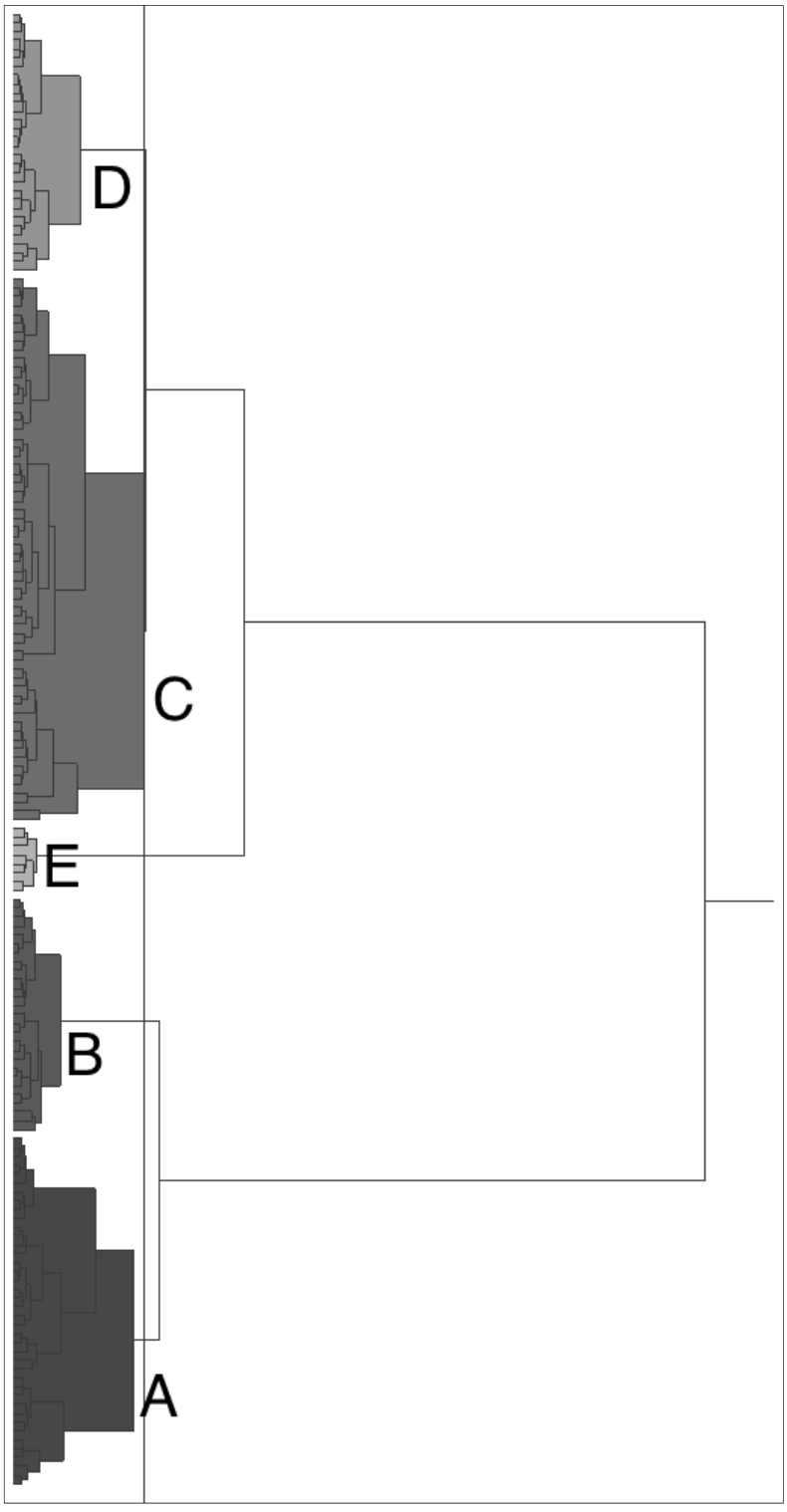


Fig. 10b. Dendrogram