Applying the Levenshtein Distance to Catalan dialects: A brief comparison of two dialectometric approaches[1]

Esteve Valls, *University of Barcelona* · John Nerbonne, *University of Groningen*
Jelena Prokic, *University of Groningen* · Martijn Wieling,
University of Groningen · Esteve Clua, *University Pompeu Fabra*
Maria-Rosa Lloret, *University of Barcelona*

Abstract. In recent years, dialectometry has gained interest among Catalan dialectologists. As a consequence, a specific dialectometric approach has been developed at the University of Barcelona, which aims at increasing the accuracy of final groupings by means of discriminating the predictable components of the language from its unpredictable ones. Another popular method to obtain dialect distances is the Levenshtein distance (LD) which has never been applied to a Catalan corpus so far. The goal of this paper is to present the results of applying the LD to a corpus of Catalan linguistic data, and to compare the results from this analysis both with the results from Barcelona and the traditional classifications of Catalan dialectology.

Keywords: dialectology, dialectometry, Levenshtein distance, underlying differences, aggregate differences, Corpus Oral Dialectal.

1. INTRODUCTION

During the last decade, dialectometry has gained interest among Catalan linguists. In the mid 90's, the main handbooks of Catalan dialectology (Veny 1982) still examined linguistic variation from a traditional point of view, i.e. taking into account bundles of isoglosses to

Data de recepción: 20-04-2010 Data de aceptación: 27-07-2010.

[1] This research is sponsored by the Spanish Ministerio de Ciencia e Innovación (project HUM2007-65531: Explotación de un corpus oral dialectal (ECOD)) and the European Regional Development Fund. It also profits from an FPU grant (Formación del Profesorado Universitario).

divide the Catalan domain into several dialect areas. This lack of an aggregate perspective, in addition to the fact that most descriptions were based on old data, stimulated linguists to design a new corpus of contemporary Catalan: the *Corpus Oral Dialectal* (COD). This corpus (Viaplana et al. 2007) has been built and systematized at the University of Barcelona, and contains more than 660.000 phonetic and morphological items gathered from the 82 county towns of the whole Catalan-speaking area. The questionnaire was designed with computerization in mind[2].

Unlike some other methods, which are item centered and superficially oriented (that is, based on phonetic outputs), the dialectometrization of the COD data has been done from a different perspective (Clua & Lloret 2006). Basically, the Barcelona approach proposes to capture the differences among varieties not only quantitatively but also qualitatively, by means of analyzing the underlying differences that would remain invisible in the phonetic data. The aim is to increase the accuracy of the final groupings. This method was first introduced by Viaplana (1999) and is a distinctive feature of Catalan dialectometry methods.

Despite the growing interest in dialectometry among Catalan linguists, the Levenshtein distance (LD) is still virtually unknown and has never been applied to Catalan language yet. The main purpose of this paper is to show the first results of applying the LD to the COD data, and to compare the dialect groupings that arise to the non-LD Barcelona approach (Clua et al. 2008).

In addition, we will also take advantage of several mapping techniques available in the L04 package[3], such as noisy clustering, to detect stable clusters in the two approaches, and multidimensional scaling to visualize dialectometric distances. These techniques have never been used in Catalan dialectometry until now.

In the following, we will give a brief overview of the studies that have contributed to the emergence of a Catalan dialectometrical approach (Section 2) and we introduce the main characteristics of the *Corpus Oral Dialectal* (Section 3). Subsequently, we present two different dialectometric approaches to examine linguistic variation: first, we give some examples of the kind of linguistic analysis we pursue to distinguish regular phonetic facts from underlying differences, and we evaluate the consequences of such distinctions for dialectometry (Section 4.1); second, we give a basic explanation about the Levenshtein distance (Section 4.2). The results using both distance measurements are discussed in Section 5, after which the differences are discussed in Section 6. Finally, we discuss some of the advantages and drawbacks of both methods and we mention some of the prospects for future research in this field (Section 7).

---

2 More information about the COD can be found in the following website: www.ub.edu/lincat.

3 More information about this software package for dialectometry and cartography, developed by Peter Kleiweg, is available at http://www.let.rug.nl/~kleiweg/L04/.

2. THE RISE OF A CATALAN DIALECTOMETRY: FROM SÉGUY TO THE *CORPUS ORAL DIALECTAL*

The connection between dialectometry and the Catalan language dates back to the inaugural study of Séguy (1971). In that paper, Séguy applied the Hamming distance to nine linguistic atlases. Two of them, the *Atlas Lingüístic de Catalunya* (ALC) and the *Atlas Linguistique des Pyrénées Orientales* (ALPO) contained exclusively Catalan data, whereas a third atlas, the *Atlas Linguistique de la France* (ALF), included some data in this language, collected in France at the eastern edge of the French-Spanish border.

This initial use of Catalan in dialectometric studies was enhanced by Enric Guiter (Sardà & Guiter 1975, Guiter 1978). This scholar, who is considered to be one of the fathers of the emerging methodology, was born in the Catalan speaking department of France, and devoted a number of years to apply dialectometry to several Catalan corpora, like the ALC and the ALPO. His work had an immediate impact both in local and international levels: on the one hand, his recommendation of using a minimum amount of a hundred maps (or glosses) to get reliable results was followed by several linguists, such as Goebl (1993) in the so-called Dialectometric School of Salzburg. On the other hand, Guiter succeeded in capturing the interest of a few Catalan dialectologists who published dialectometric works during the 70's. Some of these primary results can be found in Sardà & Guiter (1975), Costa (1983) and Guiter (1978).

In spite of this initial success, only a couple of remarkable works using dialectometry appeared during the next two decades. Polanco (1992) applied the methods developed by Guiter and Goebl to another corpus partially built with Catalan data: the *Atlas Lingüístico de la Península Ibérica* (ALPI). Later on, Ortega (1998) made use of the hierarchical cluster analysis for the first time in Catalan dialectology.

All these studies were attempts to incorporate more modern and objective techniques to the study of Catalan dialects from an aggregate perspective. However, no efforts were made to create a stable group of linguists focussed on exploiting the possibilities offered by the dialectometric methods available. As a consequence, these early works were unconnected and contributed to the development of Catalan dialectometry in an isolated way.

This isolation was finally overcome thanks to the creation and systematization of the COD. Not only is it the most important corpus dealing with actual Catalan dialects, but it is also the point of departure of a new methodology that has been used, with some minor changes, in all recent papers where dialectometric techniques have been applied to the COD data. The pioneer using this new approach, which we introduce in Section 4.1, was Viaplana (1999). He also showed the first results of applying hierarchical cluster analysis to the whole Catalan-speaking area, as well as some other mapping techniques, such as the additive trees developed by Sattath and Tversky (1977). Recently, many other papers have followed the guidelines established by Viaplana: for example, Clua (1998, 1999 and 2004) who proposes a classification of the

Valencian dialects; Clua & Lloret (2006), where the main tenets of the methodology set by Viaplana are presented and discussed; Clua (2010), who shows the first results of applying dialectometry to a subset of data extracted from the COD; and finally, Clua et al. (2008), who examine for the first time the global dendrograms from the complete dialectometrization of the COD data. These last results are of great interest in the current paper, as one of our major goals is to compare these final dendrograms with the ones obtained via the LD.

3. DATASET: THE *CORPUS ORAL DIALECTAL*

The COD is a corpus of contemporary Catalan that has been gathered and systematized since 1991 at the Catalan Philology Department of the University of Barcelona. Data were collected with computerization in mind through a questionnaire of approximately 600 phonetic and morphological items and recordings of ten-minute long samples of informal speech. Several fieldworkers interviewed 2 or 3 speakers in each of the 82 county towns of the whole Catalan-speaking area, which nowadays is divided into 4 different states: Spain, France, Andorra and Italy, where Catalan is spoken in the city of l'Alguer, in Sardinia (see Figures 1 and 2).

The informants had to adhere to the following criteria: they had to be descendants of parents born in the same locality; they had to be aged 30-45 (so that they had not been taught in Catalan at school); they had to be middle class citizens and have a minimum amount of formal education (no more than primary school); and they had to have lived in the same village all their life. The reason for selecting these localities and speakers was to build a representative corpus of the Catalan spoken by the majority of the population, as most inhabitants of the Catalan domain are settled in urban areas. As a consequence, the COD differs from previous surveys in the sense that it does not reflect the most conservative varieties of some hidden rural areas, but the dialects spoken by large masses of the Catalan population.

Figure 1. Context map of the Catalan-speaking area, including some important cities: (*1*) *Perpinyà* (*France*); (*2*) *Andorra la Vella* (*Andorra*); (*3*) *Girona*, (*4*) *Barcelona*, (*5*) *Tarragona*, (*6*) *Lleida* (*Autonomous Community of Catalonia*, *Spain*); (*7*) *Fraga* (*Autonomous Community of Aragon*, *Spain*); (*8*) *Castelló de la Plana*, (*9*) *València*, (*10*) *Alacant* (*Autonomous Community of the Valencian Country*, *Spain*); (*11*) *Eivissa*, (*12*) *Palma*, (*13*) *Maó* (*Autonomous Community of the Balearic Islands*, *Spain*); *and* (*14*) *l'Alguer* (*Sardinia*, *Italy*).

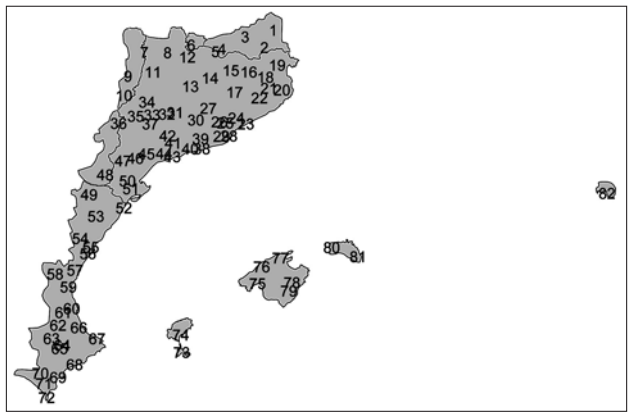


Figure 2. The 82 county towns studied in the COD (*see Table 5 at the end of the paper*).

The data have been phonetically transcribed and systematized in several databases that contain 135,480 phonetic items and 532,508 morphological items. At the same time, a smaller dataset was selected for dialectometric purposes. This dataset includes 356 glosses per locality (majority forms were selected in cases with more than one output) and it contains up to 29.192 items, corresponding to: verbs (20.500 items), articles (1.312 items), possessive pronouns (1.968 items), clitic pronouns (4.264 items), personal pronouns (656 items), demonstrative pronouns (246 items) and locative adverbs (246 items). Both the results presented in Clua et al. (2008) and those obtained via LD are based on this dataset.

4. TWO APPROACHES TO MEASURE THE LINGUISTIC DISTANCE

4.1. Recent Catalan dialectometry: a system to account for phonological differences

The results published in Clua et al. (2008), as well as the rest of dialectometric papers based on the COD so far, proceed from the assumption that a linguistic analysis of the data must be carried out before applying the measure of distance. From this point of view, it is crucial to discriminate the unpredictable components of the language (i.e. the underlying morphological differences) from its predictable elements (i.e. the regular phonological phenomena that produce the phonetic outputs). As illustrated by the following example, this distinction allows taking into consideration some structural differences that would remain invisible in the surface forms. We illustrate this point by looking at the complete paradigm of the first person singular pronominal clitic in Valencian Catalan (it appears in bold in the examples below); the first two forms are proclitics followed by a consonant (a) and a vowel (b); the last two forms are enclitics preceded by a consonant (c) and a vowel (d):

(1)		Variety 1	Variety 2	Variety 3	
	a. **em** rente	[**em**rén̪te]	[**me**rén̪te]	[**me**rén̪te]	"I wash myself"
	b. **m'**escolta	[**m**eskól̪ta]	[**m**eskól̪ta]	[**m**eskól̪ta]	"he listens to me"
	c. escoltant**-me**	[eskol̪tám**me**]	[eskol̪tám**me**]	[eskol̪tám**me**]	"listening to me"
	d. renta**'m**	[rén̪ta**m**]	[rén̪ta**m**]	[rén̪ta**me**]	"wash me"

The examples in (1) show that all the varieties display a non-syllabic form [m] and different syllabic forms: one with a vowel before the consonant ([em], row a) and another with the vowel after the consonant ([me], row c, for example). In variety 1, [em] appears before a verb that begins with a consonant (1a), while [me] appears after a verb that ends in a consonant (1c). In variety 2, [me] shows up in these two cases (1a, c). In variety 3, [me] further appears after a verb ending in a vowel (1a, c, d). From the point of view of traditional approaches, the linguistic distance between these three varieties is very similar. All the varieties show the same forms in (1b) and (1c). Varieties 1 and 3 differ in two cases: (1a) and (1d). Variety 2 differs in one form with respect to variety 1, (1a), and in one other form with respect to variety 3, (1d).

From the point of view of surface approaches, the distance between varieties 1 and 2 has the value 1 because they differ in one form only: [emrén̪te] vs. [merén̪te]. Varieties 1 and 3 show a linguistic distance of 2 because they differ in two forms: [emrén̪te] vs. [merén̪te] and [rén̪tam] vs. [rén̪tame]. Varieties 2 and 3 display a distance of 1 because they differ in one form: [rén̪tam] vs. [rén̪tame]. This is illustrated in Table 1.

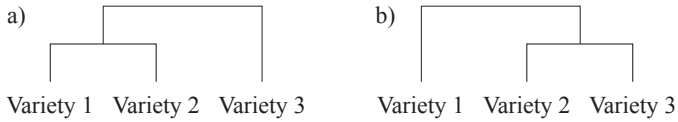
Table 1. Similarity matrix based on phonetic data

	Variety 1	Variety 2	Variety 3
Variety 1			
Variety 2	1		
Variety 3	2	1	

In a dendrogram based on this data, either varieties 1 and 2 are grouped closer than 3, because they differ in one form only (see Figure 3a), or varieties 2 and 3 are grouped closer than 1, because they also differ in one form only (see Figure 3b).

This situation, however, is notably different if we analyse the same data on the basis of syllabification, that is, distinguishing the underlying differences from the ones that are due to regular phenomena. In fact, the aim of the dialectometric approach developed in Barcelona is to calculate the linguistic distance by adding: (1) the differences among the underlying morphological forms and (2) the differences among the phonological processes that transform them into the phonetic outputs. As a consequence, the first step was to determine: (1) a morphological underlying form for every single item of the corpus; and (2) which phonological rules operated on each morphological underlying form.

Figure 3. Dendrograms based on the phonetic data of Table 1



Under this view (see (2) below), varieties 1 and 2 have a single underlying form (/m/) and the vowel [e] is inserted in order to satisfy syllabic requirements (for example, by means of the following rule in variety 1, row a: Ø → e / __m#C and the following rule in variety 1, row c: Ø → e / C#m__). That is, as in other contexts, epenthesis applies when the addition of the clitic to the verb creates a sequence that cannot be properly syllabified. The difference between these two dialects lies in the position of the epenthesis. In variety 1, the epenthetic vowel always appears at the periphery of verb-clitic sequences, i.e. at the beginning in (2a) but at the end in (2c). In variety 2, instead, it is always placed to the right of the clitic, i.e. [me] in (2a) and (2c). Variety 3 is completely different. The crucial example here is the last one, i.e. *renta'm* [rén̪ta**me**] (2d). In this case, the verb ends in a vowel and, thus, there is no syllabic reason to assume that the vowel of the clitic is inserted through epenthesis to repair syllabification. For this variety, it is more coherent to establish that the underlying form of the clitic contains the vowel (/me/), although this vowel deletes when it appears in contact with another vowel (e → Ø / m__#V), cf. *m'escolta* [**m**eskól̪ta] in (2b). This vowel also deletes in other vocalic contexts in the language (cf. *entr**e** **a**mics: entr[a]mics* 'between friends', ***no és** tan gran: n[o]s tan gran* 'it is not that big').

(2)		Variety 1 /**m**/	Variety 2 /**m**/	Variety 3 /**me**/	
a. **em** rente	[**em**rén̪te]	[**me**rén̪te]	[**me**rén̪te]	"I wash myself"
b. **m**'escolta	[**m**eskól̪ta]	[**m**eskól̪ta]	[**m**eskól̪ta]	"he listens to me"
c. escoltant-**me**	[eskol̪tám**me**]	[eskol̪tám**me**]	[eskol̪tám**me**]	"listening to me"
d. renta'**m**	[rén̪ta**m**]	[rén̪ta**m**]	[rén̪ta**me**]	"wash me"

In other words, variety 3 has preserved the old shape of the clitic (/me/, from the Latin form *me*), but in certain contexts the vowel deletes in accordance with the regular phonology of the language. Unlike variety 3, varieties 1 and 2 have re-structured their system. They show a single-consonant underlying form (/m/) that is accompanied by epenthesis for syllabic reasons. Therefore, the linguistic distance between varieties 1 and 2 is indeed smaller than that with variety 3, which has a different underlying representation. We will next show how our analysis tries to capture this fact.

The similarity matrix presented in Table 2 shows that, as for *morphological underlying differences* concerning the four forms under study (that is, the complete paradigm of the

first person singular pronominal clitic in Valencian Catalan), varieties 1 and 2 have zero differences (both have the same /m/ underlying form), but variety 1 with respect to 3, and 2 with respect to 3 show 4 differences (variety 3 departs from a /me/ underlying form). We count each position in the paradigm as one point of difference.

Table 2. Similarity matrix based on the phonological analysis (I). Morphological underlying differences: /m/$_{1,2}$ vs. /me/$_{3}$			
	Variety 1	Variety 2	Variety 3
Variety 1			
Variety 2	0		
Variety 3	4	4	

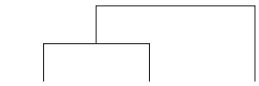
The similarity matrix presented in Table 3 further calculates the differences concerning the *phenomena involved*. Here, varieties 1 and 2 differ only in the position of the epenthesis (2a). Varieties 1 and 3 and varieties 2 and 3 differ in displaying or not epenthesis (2a and 2c) and vowel deletion (2b).

Table 3. Similarity matrix based on the phonological analysis (II). Differences in the phenomena involved			
	Variety 1	Variety 2	Variety 3
Variety 1			
Variety 2	1		
Variety 3	3	3	

The next step consists of adding the values contained in Table 2 (concerning the morphological underlying differences) and the values contained in Table 3 (concerning the phonological phenomena involved). In accordance with our analysis (see Table 4), the resulting dendrogram (Figure 4), shows a closer relation between varieties 1 and 2, and, significantly, a larger distance between these two and variety 3.

Table 4. Similarity matrix based on phonological data			
	Variety 1	Variety 2	Variety 3
Variety 1			
Variety 2	1		
Variety 3	7	7	

Figure 4. Dendrogram based on the phonological analysis of Table 4



So far, we have tried to demonstrate the importance of such an analysis to capture structural differences that would otherwise remain invisible, and we have also examined some of its consequences on the final classification of the varieties. This procedure is indeed the keystone of the dialectometric approach developed around the COD, as the linguistic distance between varieties is the result of applying a measure of distance to two different databases: one containing the underlying morphological forms and another one comprising the approximately 60 phonological phenomena involved. To simplify the analysis, each phenomenon has been assigned a number from 1 to 60. In addition, all the inputs (both the underlying forms and the processes) have been manually aligned in a previous step. Next we introduce the measure of distance:

$$dist(i,j) = \frac{\sum_{k=1}^{long} dif_k(i,j)}{long} \times 100$$

That is, the linguistic distance between two varieties (i, j) is the result of the summation of their differences (each having a value of 1) with regard to a linguistic variable k and dividing them by *long*, which is the length of each item compared. It is fundamental to take into account that the variables can be either the phonemes that make up the underlying forms or the phonological rules that transform them into the phonetic outputs.

As a result of this process we obtained a distance matrix, to which we applied the mapping techniques introduced in Section 5.

4.2. The Levenshtein distance: measuring the phonetic distance

The Levenshtein distance (Levenshtein 1966, also known as edit distance) is a string comparison procedure that calculates the phonetic distance between two phonetic strings. To obtain this distance, the Levenshtein algorithm seeks the least costly set of basic operations (insertions, deletions and substitutions) needed to transform one string into another. In the simplest version of the algorithm, these three operations have the same cost, as can be seen in the example below, based on two pronunciations of a conjugated form of the Catalan verb *servir* 'to serve' (specifically *servís* '(if I) served'). In this case, the final distance between the two pronunciations is 3:

(3)				
	Variety 1	s e ɾ β i s k é s	delete s	1
		s e ɾ β i k é s	substitute k/ɣ	1
		s e ɾ β i ɣ é s	insert ɛ	1
	Variety 2	s e ɾ β i ɣ é s ɛ		
	Total			3

From a different perspective, the procedure can also be seen as the result of aligning two strings of phonetic segments. In these alignments, phonetic overlap is binary, so that non-identical phones contribute to phonetic distance, whereas identical ones do not. In order to increase accuracy, we used a common modification of the Levenshtein algorithm, allowing only alignments of consonants with consonants and vowels with vowels. In addition, no length normalization has been applied, as this has been found to give the best results in dialectometric analyses (Heeringa et al. 2006). The following example illustrates the alignment of the two pronunciations compared in (3):

(4)

Variety 1	s	e	ɾ	β	i	s	k	é	s	
Variety 2	s	e	ɾ	β	i		ɣ	é	s	ɛ
Total						1	1			1

The use of the Levenshtein distance to calculate the linguistic distance between varieties was first introduced by Kessler (1995), who applied it to several Irish dialects. Since then, it has been used to analyze linguistic variation in more than ten other languages, such as German (Nerbonne 2010), Dutch (Heeringa 2004), Frisian (Heeringa 2005), Bulgarian (Osenova et al. 2010) and the languages of Gabon (Alewijnse et al. 2007). Furthermore, the procedure has also been successfully applied to other sorts of research: see, for example, Heeringa & Gooskens (2003), who examine 15 Norwegian dialects perceptually and acoustically; or Gooskens & Bezooijen (2006), where they compare the level of mutual comprehensibility between Afrikaans and Dutch.

Despite many studies that have used the Levenshtein distance during the past thirteen years, no study applied the LD to a Catalan corpus before the current study. To use the LD, the phonetic symbols had to be converted in X-SAMPA to produce a machine-readable phonetic transcription for the L04 package. After applying the LD to the corpus, a resulting distance matrix was obtained. In the next section we introduce the techniques used to visualize the linguistic distances contained in the distance matrices based on the LD and the Barcelona approach.

We are also interested in the differences between the BCN and LD methods of assessing aggregate linguistic distance. LD works on concrete pronunciations (phonetic transcriptions), aligning them automatically, and without first subjecting them to phonological analysis,

which is part of the BCN approach. This means that LD is relatively straightforward to apply, for example due to the automatic alignment, which must be supplied manually in the BCN approach. The BCN approach assesses differences with respect to phonological analyses, which are more succinct characterizations of varieties, but they are also subject to variation, as different analysts may propose different underlying forms. LD clearly has the potential disadvantage of being insensitive to underlying differences which emerge in the BCN phonological analysis step. We undertake the current study in an attempt to understand the relative importance of these advantages and disadvantages. In general we expect LD to be more sensitive to superficial (concrete) differences and for BCN to be sensitive to the underlying distinctions made in the phonological analysis step. Naturally there will be overlap.

5. TECHNIQUES TO VISUALIZE AGGREGATE DIFFERENCES

When Séguy presented the first dialectometric approach in his paper of 1971, he argued two main reasons for using it: first, he wanted to overcome the subjectivity of the traditional analysis, based on bundles of isoglosses; second, he intended to make use of all the data contained in the linguistic atlases available at the moment, as they had historically been underexploited. But there was a third crucial goal in the early studies of Séguy: the intention to improve the mapping techniques to visualize the results of his analysis. This is the reason why he introduced the first network maps ever used in dialectology: these maps, which can be found in the sixth and last volume of the *Atlas Linguistique de la Gascogne*, show the Hamming distances between all the localities of the studied language area (Séguy 1973). They are, therefore, the first attempt to present linguistic distances in a more comprehensible and accessible way.

Since then, the interest for developing proper techniques to map aggregate variation has become a pillar in all dialectometric approaches. In Salzburg, Goebl and his team have created a wide range of maps to give a heuristically comprehensive answer to the question of how a local dialect deals with its geolinguistic environment. In particular, Goebl (1993) tends to point out the importance of the so-called *similarity maps*. Similarly, several mapping techniques have been successfully developed at the University of Groningen. Nerbonne (2010) gives a detailed introduction with respect to the motivation and methods used to project aggregate variation to maps.

In the following we present the first results of applying five mapping techniques to the COD data that have not been used in Catalan dialectology before. They are a direct consequence of the collaboration established between the Catalan Philology Department of the University of Barcelona and the Alfa-Informatica Department of the University of Groningen a few months ago, and will allow us to examine the relations between Catalan dialects more accurately.

5.1. Network maps

The "network maps" (also called "link maps") are simple initial visualizations of the aggregate differences made by drawing lines between data collection sites. The darkness of these lines is inversely proportional to the linguistic distance between the sites, so that lighter lines connect more linguistically distant sites (Nerbonne 2010).

These first visualizations allow us to detect, on the one hand, the linguistically more homogeneous areas and, on the other hand, the transition areas between dialects. These are strips of discontinuity, characterized by light lines between contiguous varieties. Figures 8 and 10, which constitute the first network maps built on the basis of Catalan data, reveal how linguistic variation is structured geographically in the Catalan speaking area. In Section 6 we will come back to this point.

5.2. Stable clustering: noisy clusters and composite cluster maps

As mentioned before, Clua et al. (2008) applied hierarchical agglomerative clustering to the distance matrix obtained from the COD data in order to get a global dendrogram of the Catalan varieties. Clustering is a well-known procedure to seek groups of close varieties, and has been used in dialectometry since Shaw (1974). It is an iterative procedure that selects the shortest distance in a matrix and fuses the two data points that gave rise to it. As these two points form a new cluster, the distance between it and the remaining elements in the matrix must be recalculated. In the end, it produces a hierarchically structured dendrogram as the one published in Clua et al. (2008). The clustering algorithm used was the so-called UPGMA (Unweighted Pair Group Method with Arithmetic Mean).

Although the use of regular clustering has become more and more popular among linguists interested in dialectometry, it is also broadly accepted that it lacks stability. This is due to the fact that clustering looks for the minimum distance between two points in a matrix, and sometimes several pairs of elements may show similar distances. As a consequence, small differences in the input data matrix can lead to considerably different clusters.

To overcome this instability, two main methods have been suggested and tested during the last years: noisy clustering (Kleiweg et al. 2004) and bootstrapping (Nerbonne et al. 2008). Briefly, noisy clustering can be viewed as a procedure in which different amounts of random noise are added to the distance matrix during repeated clustering. Bootstrapping consists of varying the input dataset in several clustering iterations, allowing some words to be repeated. The result of both techniques is a consensus (or composite) dendrogram.

To interpret a composite dendrogram, two facts must be taken into account: first, the numbers associated with the brackets indicate how many times these varieties have clustered together in the iterations of the process; second, the length of the brackets is the mean cophenetic distance found in the runs where this group emerged.

Moreover, it is possible to project and visualize this information in a "composite cluster map". This mapping technique starts by dividing the studied area according to the Voronoi polygonisation, so that lines are drawn between adjacent sites. Note that these divisions are also used to get the similarity maps developed in Salzburg (see Goebl 1993). Afterwards, these borders are shaded in a way that their darkness is directly proportional to the cophenetic distances between the two contiguous localities in the consensus dendrogram.

As we had both distance matrices available we used noisy clustering to obtain stable dendrograms from both approaches.

5.3. Multidimensional Scaling (MDS) and RGB maps

We also applied another statistical technique to the COD data: multidimensional scaling (MDS), which aims at reducing high dimensional data to a smaller number of dimensions. MDS was first introduced in dialectology by Black (1976), who measured the distance among several dialects of four linguistic groups, located in the Philippines, Africa and North America. There are two main reasons for using the MDS: first, because, unlike clustering, it is a stable method to plot the linguistic distances; and secondly, because it gives us the possibility to examine the relations between varieties in more detail than by using a consensus dendrogram (see Nerbonne 2010: 15).

To visualize the MDS results we use RGB maps. In these maps, the MDS results are visualized by colouring each data point red, green and blue in proportion to its first, second and third MDS coordinates, respectively. An advantage of MDS maps is that they allow interpretations in terms of dialect continua.

6. THE ANALYSIS OF THE RESULTS: DIFFERENCES BETWEEN TWO DIALECTOMETRIC APPROACHES

Traditional dialectology has historically divided the Catalan language into two main dialect areas: Eastern Catalan and Western Catalan. In turn, a number of subdialects are located on both sides of this main border: Northern, Central, Balearic and Algherese Catalan are considered to belong to the Eastern Catalan group; similarly, North-Western and Valencian Catalan are the two major varieties of the Western group. At the same time, dialectologists have partitioned the Valencian dialect into three varieties, namely Northern Valencian, *Apitxat* and Southern Valencian. Finally, some linguists also claim the existence of a transition variety between North-Western and Valencian Catalan: the so-called *Tortosí*.

If we now examine the results obtained via the two dialectometric methods presented above, the first remarkable fact is the number of similarities they share. To begin, the linguistic distance tables correlate very highly (r= 0.868). As Figures 7 to 10 show (especially the two composite cluster maps), both approaches succeed in identifying several well-known facts

in Catalan dialectology: (1) they both trace exactly the same border between Eastern and Western Catalan; (2) they both confirm the fact that Central Catalan is the most homogeneous subdialect and has no major internal divisions; (3) they both highlight that Benavarri and l'Alguer are isolated localities in their respective contexts; (4) they both draw a border to the south of Vinaròs (this line is darker using the LD though); (5) they both point out the idiosyncrasy of the Balearic varieties; and (6) they both display a homogeneous dialect area corresponding to the big *plateau* of *Lleidatà*, the most spoken North-Western subdialect. This situation reproduces to some extent the traditional divisions (see Veny 1982), but it is especially relevant because it fully agrees with the first dialectometric analysis exclusively focussed on the North-Western varieties (see Viaplana 1999). To be precise, the borders of *Lleidatà* according to both methods are exactly coincident with those that separate the most conservative North-Western varieties from those more influenced by the Standard Catalan, which have partially lost some of its North-Western distinctive features. Additionally, in the consensus dendrogram obtained via the LD, the locality of Benavarri clusters together (at a remarkable distance, though) with the rest of the North-Western Catalan varieties.

Despite this overall resemblance, several differences arise when we go through the results in more detail. The most striking difference in the groups detected by the two techniques concerns the location of the Valencian varieties. While LD groups the Valencian area together with the rest of the Western varieties, in accordance with the traditional classifications (Veny 1982), the other consensus dendrogram analyses most of these varieties as a separate, most important group; only four localities (those corresponding to Northern Valencian: Albocàsser, l'Alcora, Morella and Castelló de la Plana) are included in the North-Western cluster, giving evidence of the continuum that spreads between the Autonomous Communities of Catalonia and the Valencian Country. In both the LD and the BCN approach, the splits are made very confidently.

Although we cannot give a definitive explanation for this difference, we know that: (1) the morphological features of Valencian are the most peculiar among Catalan dialects; (2) 2/3 of the corpus are based on verbal morphology; and (3) the BCN approach tends to increase the weight of the morphological differences, as we saw in Section 4.1. These three facts might be to some extent responsible for the different classification of the Valencian varieties in the results.

The next difference concerns the Valencian varieties internally; as we see in Figure 5, the consensus dendrogram obtained using the approach from Barcelona identifies two main groups of Valencian varieties: a group encompassing the Central and Southern varieties (from Alacant to Guardamar) and a group containing the *Apitxat* varieties (Alzira, Sagunt, Llíria and València). In addition to Vinaròs, which clearly falls in the *Tortosí* area, the other varieties that belong to the Autonomous Community of the Valencian Country (Albocàsser, l'Alcora, Morella and Castelló de la Plana) are clustered in the North-Western group. It is a remarkable fact, since the classification of these varieties has always been a matter of

discussion (see Pradilla 2009: 121-140); some authors believe they share so many features with the *Tortosí* that they should be considered a transition area between the dialects spoken in Catalonia and the ones spoken in the Valencian Country; some others, instead, have claimed the convenience of considering them a separate subdialect (the so-called *Northern Valencian*).

If we turn to the results obtained via LD (see Figure 6), we see that all Valencian varieties are clustered in only one global group, within which no reliable subdialects emerge –not even the Northern Valencian, whose varieties amalgamate with the rest of the Valencian localities. Moreover, this dendrogram does not identify the *Apitxat* subdialect, and gives evidence of the fact that the linguistic analysis of the data must be responsible for its different classification in the previous consensus dendrogram. We will return to this point later.

The third remarkable difference regards the group of varieties traditionally defined as Northern Catalan (i.e., those located in the *French Dépertement des Pyrénées Orientales*: Prada, Sallagosa, Perpinyà and Ceret). Whereas in Figure 5 three of them cluster together, in Figure 6 they do not form a homogeneous group anymore and cluster instead with the Eastern varieties in an isolated way. It is worthwhile having a look at Figures 12 and 14 to better appreciate this difference. We suspect that these groupings might reflect the difficulties in meeting reliable informants in the area (as only a few people use Catalan in their everiday life in France) rather than the dialect spoken there. That's why we would like to revise the data collected in the *Département des Pyrénées Orientales* in the future.

We also find a fourth significant difference when we examine the relations among Balearic varieties, specifically the Minorcan ones. Although Maó and Ciutadella always form a separate group, in Figure 5 they integrate in the cluster of Balearic varieties, whereas using the LD they are not placed in this cluster anymore. On the contrary, the consensus dendrogram indicates that they can only be assigned reliably to the Eastern Catalan cluster. Although it has been noted in the past that the Minorcan way of speaking is the closest one to the Central Catalan amongst all Balearic varieties, we also have to take into account that the corpus used does not include phonosyntactic data, where Minorcan varieties show a lot of common solutions with the other Balearic subdialects.

Finally, the BCN approach rather favors the isolation of Benavarri, a locality belonging to the Autonomous Community of Aragon where a transition variety that shares features with Aragonese is spoken.

So far, we have mentioned some major differences between the two approaches on the basis of the specific location of a group of dialects with respect to the rest of groupings. However, there are still a couple of remarkable divergences that arise when we carefully observe the MDS plots and the RGB maps (see Figures 11, 12, 13 and 14; the correlation between the two methods is $r = 0.868$). The first idea we extract from the comparison of Figures 11 and 13 is that the dialectometric approach from Barcelona seems to favor the emergence of some more homogeneous clusters that are linguistically relevant (see, for

example, the *Apitxat* varieties, the North-Valencian varieties or the Balearic varieties). We hypothesize that it is due to the fact that it detects partial similarities where the LD only counts differences. Secondly, both the MDS plots and the RGB maps display a division between Eastern and Western varieties which seems to be a bit sharper using the LD. In what follows we will try to give some examples that can account for these differences.

(5)

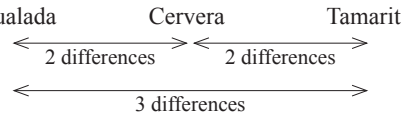
	Central Catalan	North-Western Catalan		
	Igualada	Cervera	Tamarit de Llitera	
(1) em rento	[əmrén̪tu]	[emrén̪to]	[merén̪to]	"I wash myself"
(2) escoltant-me	[əskul̪támmə]	[askol̪támme]	[askol̪támme]	"listening to me"

If we take again into consideration the first singular pronominal clitic, we can observe several differences between the varieties of Igualada (located in the Eastern Catalan area), Cervera (a North-Western variety located beside the border between Eastern and Western Catalan) and Tamarit de Llitera (one of the most conservative North-Western dialects). Using the LD, these differences would result in the following distances:

1. Between Igualada and Cervera: there is 1 substitution (ə > e) in (1) plus 1 substitution (again ə > e) in (2). The result is a distance of 2.
2. Between Cervera and Tamarit: there are 2 differences in (1), 1 deletion (e > Ø) and 1 insertion (Ø > e). The result is a distance of 2.
3. Between Igualada and Tamarit: there are again 2 differences in (1), 1 deletion (ə > Ø) and 1 insertion (Ø > e). In addition, there is one substitution (ə > e) in (2). The result is a distance of 3.

These distances can be represented by the following diagram:

(6)



As can be observed directly, this situation changes when a linguistic analysis of the data is first carried out. According to the approach from Barcelona, the resulting distance would be as follows:

1. As the three varieties share the same single-consonant underlying form /m/, the database containing the *morphological data* counts 1 similarity between Igualada and Cervera, another one between Cervera and Tamarit and another similarity between Igualada and Tamarit.
2. When it comes to the *phonological processes* involved to insert the epenthesis, they result in the distances below:

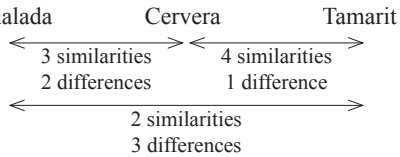
a) Between Igualada and Cervera: there is 1 similarity (insert epenthesis to the left of the nucleus) and 1 difference (the timbre of the epenthesis) in (1). There is also 1 similarity (insert epenthesis to the right of the nucleus) and 1 difference (the timbre of the epenthesis) in (2). This results in 2 similarities plus 2 differences.

b) Between Cervera and Tamarit: there is 1 similarity (the timbre of the epenthesis) and 1 difference (the epenthesis position) in (1), plus 2 similarities (both the epenthesis timbre and position) in (2). It results in 3 similarities and 1 difference.

c) Between Igualada and Tamarit: there are 2 differences (the epenthesis timbre and position) in (1) and 1 similarity (the epenthesis position) and 1 difference (its timbre) in (2). It results in 1 similarity and 3 differences.

Again, the number of coincidences/divergences obtained by adding the similarities and differences from the morphological data (1) and the phonological processes (2) can be represented in a diagram:

(7)



Although the examples above are only based on a couple of items, the diagrams are useful to display the influence of both methods in the final results. This influence shows up clearly in the MDS plots: in Figure 13, for instance, Cervera is the closest North-Western locality to the Eastern cluster, at a significant distance from Tamarit; in Figure 11, instead, both the distance between Cervera and Tamarit and the distance of the North-Western and Eastern clusters diminish. Thus, the examples seem to confirm the hypothesis that the approach used in Barcelona tends to favor higher homogeneity within some clusters and lesser distance between close clusters. Applying the LD, instead, results in a different overall pattern, with some more distance between the mainland Eastern varieties and the North-Western cluster; at the same time, the *Tortosí* clearly plays the role of a transition area between North-Western Catalan and Northern Valencian, covering the larger distance that separates these two clusters.

The last example we want to mention focuses on the *Apitxat* varieties, as their location in the two sets of results differs notably. As we mentioned, the *Apitxat* has traditionally been considered a separate subdialect because its phonological inventory lacks voiced fricatives and affricates (this is also a distinctive feature of the Benavarri area). However, we have already mentioned that the consensus dendrogram obtained via LD does not show a separate cluster with these varieties: we can see in Figure 13 that València, Llíria, Alzira and Sagunt are closely located inside the general cluster of the Valencian varieties.

In the dendrogram obtained using the approach from Barcelona, instead, these four localities form a very separate group. This difference must be probably seen as a direct consequence of the previous linguistic analysis of the data, as we illustrate next. For example, using the LD, the final distance between the varieties of Alcoi (Central Valencian) and València (*Apitxat*) with respect to the feminine plural article *les* is 0:

(8) Alcoi [le**s**]
València [le**s**] } Total distance (LD) = **0**

Using the approach from Barcelona, instead, the final distance between these two varieties is 2. On the one hand, it is due to the fact that the phonological inventory of the *Apitxat* varieties lacks the phoneme /z/, which is considered to be the underlying form of the morpheme of plural in Catalan. In the dataset containing the morphological information, thus, we would find a first difference between varieties:

(9) Alcoi /l+a+**z**/
València /l+a+**s**/

Additionally, the fact that final obstruent devoicing is a systematic phonological process in Catalan adds a second difference between the *Apitxat* varieties and the rest of dialects in the phonological level. Indeed, this process cannot occur in *Apitxat* with regard to fricatives and affricates, as their underlying forms are already voiceless. However, it is a very common process in the rest of Catalan varieties. As a consequence, the dialectometric approach of Barcelona counts 2 differences where the LD does not count any distance:

(10)

	Morphological dataset	Phonological dataset	
Alcoi	/l+a+**z**/	Final obstruent devoicing	
València	/l+a+**s**/	-	
Total distance (BCN)	1	1	= **2**

Besides, the number of occurrences of these phonemes in final word position is quite high: for example, they appear in all plural forms of articles, possessive pronouns, personal pronouns, clitics, demonstratives and locatives, and in all singular second persons of the verbs examined, among others. The consequence is a clear increase of the differences between the final linguistic distances obtained via each dialectometric approach.

7. DISCUSSION AND PROSPECTS

In this paper, we have presented the main tenets of two dialectometric approaches developed during the last decade, and we have compared the results of applying them to the same corpus of linguistic data. For visualization purposes, we have taken advantage of several mapping techniques which have not been previously used in Catalan dialectology. They clearly improve the traditional ways of projecting aggregate variation to geography used so far in the dialectometric papers based on Catalan data.

During the analysis of the results, several differences between the two methods (and with respect to traditional classifications) have arisen, and we have tried to give plausible explanations to the most striking divergences detected in the groupings. These differences have been attributed to the fact that discriminating what is predictable and what is unpredictable in the language, as is done in Barcelona, sometimes increases and sometimes decreases the total relative distances among varieties in comparison to the LD results.

However, several other factors might be responsible for some of these differences, namely: (1) the fact that in Barcelona the measure of distance is applied to a multiple aligned corpus whereas the Levenshtein Distance works on the basis of pairwise string alignments; or (2) the different nature of the methods employed to calculate the linguistic distance (that is, the measure of distance described in Section 4.1 or the Levenshtein algorithm). Unfortunately, due to the implicit differences between the methods compared, we did not succeed in isolating these factors.

As a consequence, in future research we would like to focus on clarifying the influence of these factors. It might be possible to compare the results of the Levenshtein algorithm with the results of applying the measure of distance used in Barcelona to the same multiple aligned set of phonetic data. In addition, it might be interesting to use more sensitive phonetic distances (instead of only using the vowel-consonant distinction) in the LD, for instance by automatically generating them on the basis of the phonetic data (Wieling et al. 2009).

Acknowledgements

We are grateful to Miquel Salicrú and Sergi Civit, both professors at the Department of Statistics of the University of Barcelona, for their advice with the statistical analysis. We are also thankful to Çagri Çoltekin for his help with the data conversion into X-SAMPA, and Bart Alewijnse for his interest to improve a new interface for the L04 package. Finally, we also want to thank Charlotte Gooskens, Therese Leinonen and Sebastian Kürschner for their useful feedback on the methodology developed in Barcelona and presented here.

REFERENCES

Alewijnse, Bart; Nerbonne, John; van der Veen, Lolke & Franz Manni (2007): "A Computational Analysis of Gabon varieties", in Petya Osenova et al. (eds.): *Proceedings of the RANLP Workshop on Computational Phonology*. Workshop at the conference *Recent Advances in Natural Language Processing*. Borovetz, pp. 3-12.

ALC: Griera, Antoni (1964): *Atlas Lingüístic de Catalunya*. Sant Cugat del Vallès: Instituto Internacional de Cultura Románica.

ALF: Gilléron, Jules & Édmond Édmont (eds.) (1902-1910): *Atlas linguistique de la France*. Paris: Champion.

ALPI: Navarro Tomás, Tomás (dir.) (1962): *Atlas Lingüístico de la Península Ibérica*. Madrid: CSIC.

ALPO: Guiter, Enric (1966): *Atlas Linguistique des Pyrénées Orientales*. Paris: Centre National de la Recherche Scientifique.

Black, Paul (1976): "Multidimensional Scaling Applied to Linguistic Relationships", *Cahiers de l'Institut de Linguistique de Louvain* 3 (5-6).

Clua, Esteve (1998): *Variació i distància lingüística. Classificació dialectal del valencià a partir de la morfologia flexiva*. Barcelona: Universitat de Barcelona.

Clua, Esteve (1999): "Distància lingüística i classificació de varietats dialectals", *Caplletra. Revista internacional de filologia* 26, pp. 11-26.

Clua, Esteve (2004): "El mètode dialectomètric: aplicació de l'anàlisi multivariant a la classificació de les varietats del català", in Maria Pilar Perea (ed.): *Dialectologia i recursos informàtics*. Barcelona: Universitat de Barcelona, pp. 59-88.

Clua, Esteve & Maria-Rosa Lloret (2006): "New tendencies in geographical dialectology: The Catalan Corpus Oral Dialectal (COD)", in Jean-Pierre Y. Montreuil (ed.): *New Perspectives on Romance Linguistics*, vol. 2 (Phonetics, Phonology, and Dialectology). Amsterdam/Philadelphia: John Benjamins.

Clua, Esteve (2010): "Distancia lingüística entre los dialectos del catalán a partir de los datos del COD", in Paul Danler et al. (eds.): *Actes du XXVième Congrès International de Linguistique et de Philologie Romanes* (*Innsbruck*, *3-8 septembre 2007*), vol. 4. Berlin: Walter de Gruyter, pp. 329-339.

Clua, Esteve; Valls, Esteve & Joaquim Viaplana (2008): "Analisi dialettometrica del catalano partendo dai dati del COD. Una prima approssimazione alla gerarchia tra varietà", in Gabriele Blaikner Hohenwart et al. (eds.): *Ladinometria. Miscellanea per Hans Goebl per il 65º compleanno. Edizione multiligue*, vol. 2. Vigo di Fassa: Istituto Culturale Ladino, pp. 27-42.

Costa, Jordi (1983): "Aproximació lingüística al català de la Cerdanya", in *Primer Congrés Internacional d'Història de Puigcerdà* (*1977*). Puigcerdà: Institut d'Estudis Ceretans, pp. 207-217.

Goebl, Hans (1993): "Dialectometry: a short overview of the principles and practice of quantitative classification of linguistic atlas data", in Reinhard Köhler & Burkhard B. Rieger (eds.): *Contributions to quantitative linguistics*. Dordrecht: Kluwer, pp. 277-315.

Gooskens, Charlotte & Reneé van Bezooijen (2006): "Mutual Comprehensibility of Written Afrikaans and Dutch: Symmetrical or Asymmetrical?", *Literary and Linguistic Computing* 21 (4), pp. 543-557 (15).

Guiter, Enric (1978): "Panorama lingüístic des de Besalú", in *Annals del Patronat d'Estudis Històrics d'Olot i Comarca 1978*. Olot: PEHOC, pp. 35-48.

Heeringa, Wilbert (2004): *Measuring Dialect Pronuntiation Differences using Levenshtein Distance*. Groningen: Groningen Dissertations in Linguistics 46.

Heeringa, Wilbert & Charlotte Gooskens (2003): "Norwegian Dialects Examined Perceptually and Acoustically", *Computers and the Humanities* 37, pp. 293-315.

Heeringa, Wilbert (2005): "Dialect variation in and around Frisia: classification and relationships", *Us Wurk*: *Tydskrift foar Frisistyk* 54 (i3-4), pp. 125-167.

Heeringa, Wilbert; Kleiweg, Peter; Gooskens, Charlotte & John Nerbonne (2006): "Evaluation of String Distance Algorithms for Dialectology", in John Nerbonne & Erhard Hinrichs (eds.): *Linguistic Distances*, ACL Workshop held at ACL/COLING. Sydney Shroudsburg, PA: ACL, pp. 51-62.

Kessler, Brett (1995): "Computational Dialectology in Irish Gaelic", in *Seventh Conference of the European Chapter of the Association for Computational Linguistics* (*Dublin*, *Ireland*). San Francisco: Morgan Kaufmann Publishers, pp. 60-66.

Kleiweg, Peter; Nerbonne, John & Leonie Bosveld (2004): "Geographic Projection of Cluster Composites", in Alan Blackwell; Kim Marriott & Atsushi Shimojima (eds.): *Diagrammatic Representation and Inference*. *Diagrams 2004*. Lecture Notes in Artificial Intelligence 2980. Berlin: Springer, pp. 392-394.

Levenshtein, Vladimir I. (1966): *Binary codes capable of correcting deletions*, *insertions and reversals*. Doklady Akademii Nauk SSSR 163, pp. 845-848.

Nerbonne, John (2010): "Mapping Aggregate Variation", in Stephan Rabanus; Ronald Kehrein & Alfred Lameli (eds.): *Mapping Language*, vol. 2 within series *Language and Space*. Berlin: Mouton De Gruyter, pp. 476-495, maps 2401-2406.

Nerbonne, John; Kleiweg, Peter; Manni, Franz & Wilbert Heeringa (2008): "Projecting Dialect Distances to Geography: Bootstrap Clustering vs. Noisy Clustering", in Christine Preisach; Lars Schmidt-Thieme; Hans Burkhardt & Reinhold Decker (eds.): *Data Analysis*, *Machine Learning and Applications*. *Proceedings of the 31st Annual Meeting of the German Classification Society*. Berlin: Springer, pp. 647-654.

Ortega, Joan Carles (1998): "Aplicació de tècniques de socioestadística avançada en l'anàlisi de dades dialectals: classificació dels parlars locals de la comarca de la Marina", *Sarrià*. *Revista d'Investigació i assaig de la comarca* 0. *La Callosa d'en Sarrià*.

Osenova, Petya; Heeringa, Wilbert & John Nerbonne (2010): "A Quantitative Analysis of Bulgarian Dialect Pronunciation", *Zeitschrift für slavische Philologie* 66, pp. 425-458.

Polanco, Lluís (1992): "Llengua i dialecte: una aplicació dialectomètrica a la llengua catalana", in *Miscel·lània Sanchis Guarner III*. Barcelona: Publicacions de l'Abadia de Montserrat, pp. 5-28.

Pradilla, Miquel-Àngel (2009): *La tribu valenciana*. *Reflexions sobre la desestructuració de la comunitat lingüística*. Benicarló: Onada.

Sardà, Anna & Enric Guiter (1975): "*L'Atlas Lingüístic de Catalunya* i la fragmentació dialectal del català", *Miscellania Barcinonensia* 40. Barcelona: Ajuntament de Barcelona, pp. 93-112.

Sattath, Shmuel & Amos Tversky (1977): "Additive similarity trees", *Psychometrika* 42 (3), pp. 319-345.

Séguy, Jean (1971): "La relation entre la distance espatiale et la distance lexicale", *Revue de Linguistique Romane* 35, pp. 335-357.

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

Shaw, David (1974): "Statistical analysis of dialect boundaries", *Computers and the Humanities* 8, pp. 173-177.

Veny, Joan (1982): *Els parlars catalans* (*síntesi de dialectologia*). Palma: Editorial Moll, 2002.

Viaplana, Joaquim (1999): *Entre la dialectologia i la lingüística*. *La distància lingüística entre les varietats del català nord-occidental*. Barcelona: Publicacions de l'Abadia de Montserrat.

Viaplana, Joaquim; Lloret, Maria-Rosa; Perea, Maria-Pilar & Esteve Clua (2007): *COD*. *Corpus Oral Dialectal*. Barcelona: PPU (CD-ROM).

Wieling, Martijn; Prokić, Jelena & John Nerbonne (2009): "Evaluating the pairwise alignment of pronunciations", in Lars Borin & Piroska Lendvai (eds.): *Language technology and resources or cultural heritage*, *social sciences*, *humanities and education*, pp. 26-34.

Table 5. *Names of the 82 localities of the COD* (*see Figure 2*)

1	Perpinyà	22	Santa Coloma de Farners	43	Tarragona	64	Cocentaina
2	Ceret	23	Mataró	44	Reus	65	Alcoi
3	Prada	24	Granollers	45	Falset	66	Gandia
4	Sallagosa	25	Sabadell	46	Móra d'Ebre	67	Dénia
5	Puigcerdà	26	Terrassa	47	Gandesa	68	La Vila Joiosa
6	Andorra la Vella	27	Manresa	48	Vall-de-roures	69	Alacant
7	El Pont de Suert	28	Barcelona	49	Morella	70	Novelda
8	Sort	29	Sant Feliu de Llobregat	50	Tortosa	71	Elx
9	Benavarri	30	Igualada	51	Amposta	72	Guardamar
10	Tamarit de Llitera	31	Cervera	52	Vinaròs	73	Formentera
11	Tremp	32	Tàrrega	53	Albocàsser	74	Eivissa
12	La Seu d'Urgell	33	Mollerussa	54	L'Alcora	75	Palma
13	Solsona	34	Balaguer	55	Castelló de la Plana	76	Sóller
14	Berga	35	Lleida	56	Borriana	77	Pollença
15	Ripoll	36	Fraga	57	Sagunt	78	Manacor
16	Olot	37	Les Borges Blanques	58	Llíria	79	Felanitx
17	Vic	38	Vilanova i la Geltrú	59	València	80	Ciutadella
18	Banyoles	39	Vilafranca del Penedès	60	Sueca	81	Maó
19	Figueres	40	El Vendrell	61	Alzira	82	L'Alguer
20	La Bisbal d'Empordà	41	Valls	62	Xàtiva		
21	Girona	42	Montblanc	63	Ontinyent		

Figure 5. Consensus dendrogram obtained applying the noisy clustering to the distance matrix from BCN. Number of runs: 100. Amount of added noise: 0,33

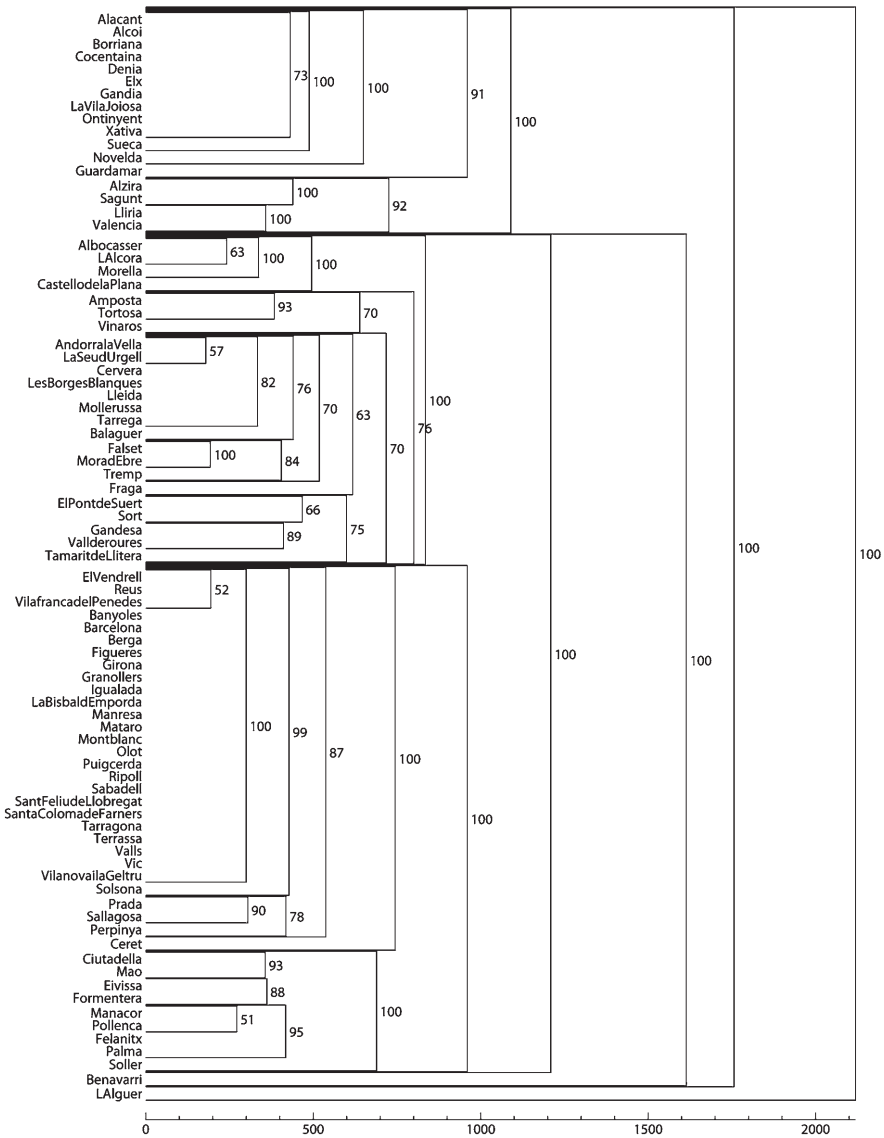
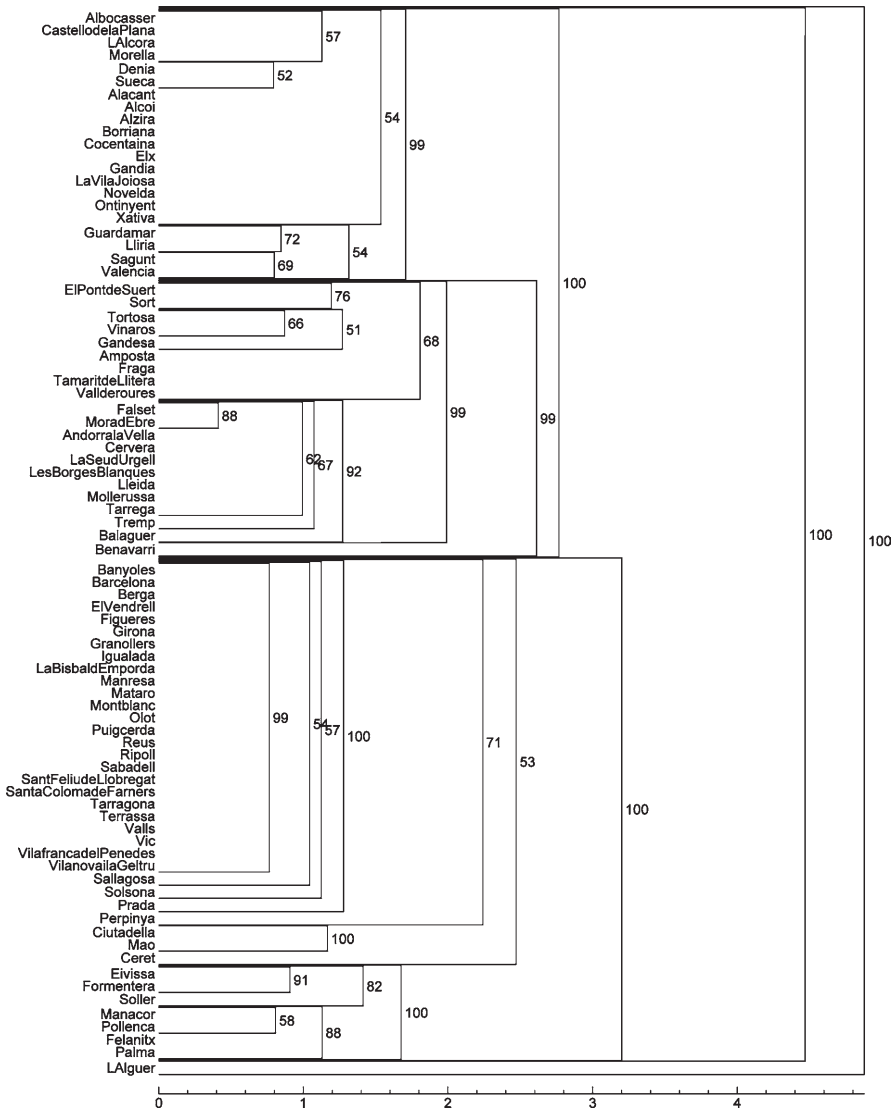


Figure 6. Consensus dendrogram obtained applying the noisy clustering to the distance matrix from the LD. Number of runs: 100. Amount of added noise: 0,33



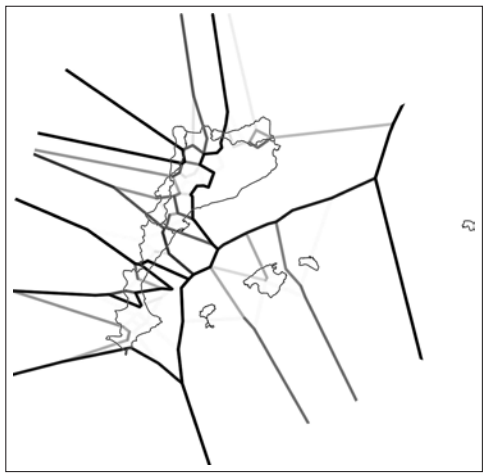


Figure 7. Composite cluster map using the approach from BCN. Number of runs: 100. Amount of noise: 1

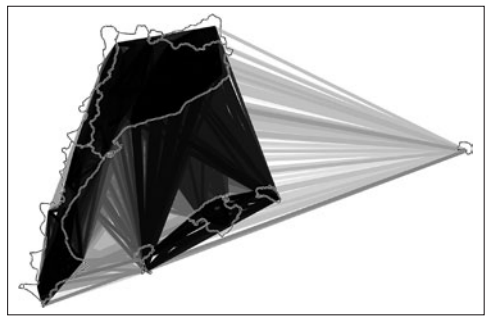


Figure 8. Network map using the approach from BCN

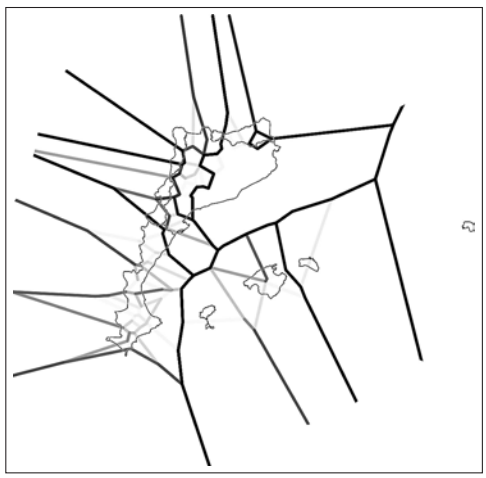


Figure 9. Composite cluster map using the LD. Number of runs: 100. Amount of noise: 1

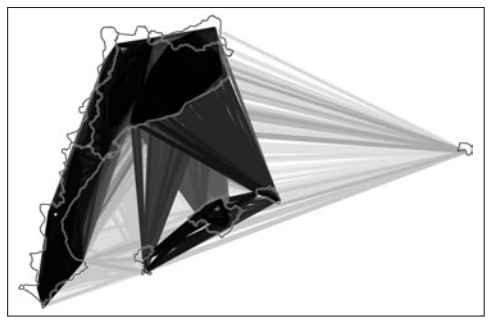


Figure 10. Network map using the LD

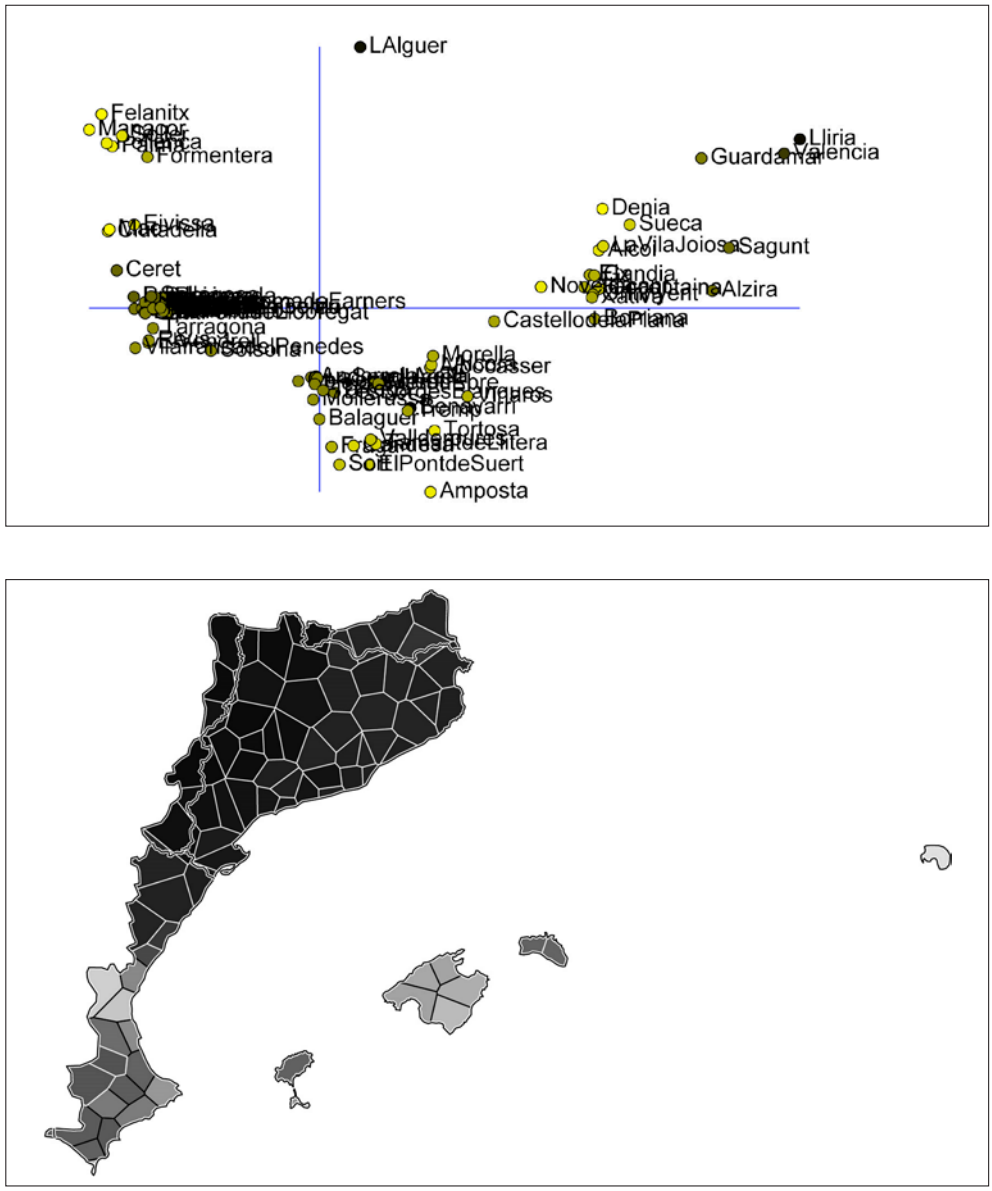
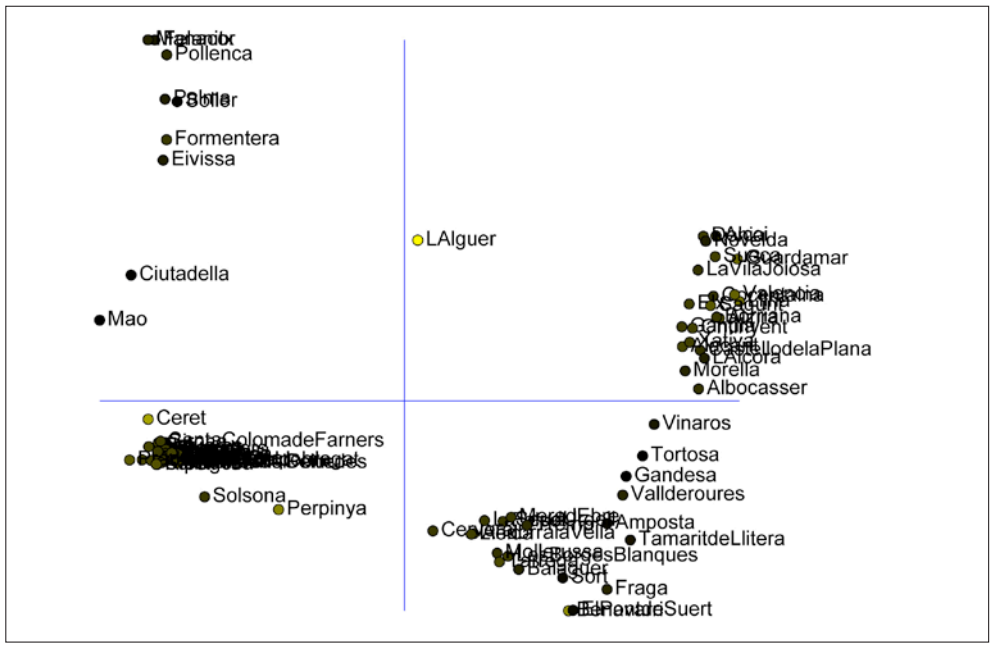


Figure 11 (*above*) *and 12* (*below*). *MDS plot and RGB map obtained using the distance matrix from BCN. R = 0,977*



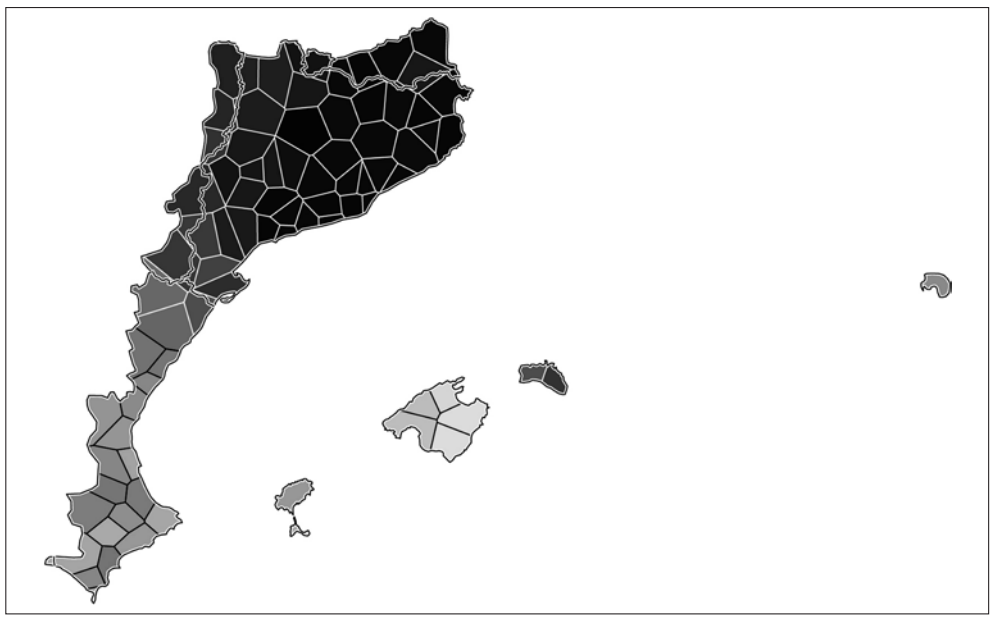


Figure 13 (*above*) *and 14* (*below*). *MDS plot and RGB map obtained using the LD. R = 0,993*