Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 198 (2015) 89 – 94

Procedia
Social and Behavioral Sciences

7th International Conference on Corpus Linguistics: Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond (CILC 2015)

# COD2: An oral dialectal corpus for the analysis of spatial and temporal variations in Catalan

Esteve Clua[a,*], Maria-Rosa Lloret[b]

[a]*Institut Interuniversitari de Lingüística Aplicada, Universitat Pompeu Fabra, Roc Boronat 138, 08018 Barcelona, Spain*
[b]*Department of Catalan Philology, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08002 Barcelona, Spain*

**Abstract**

In this article we present the most relevant new aspects from the COD2 project, which is based on an oral dialectal corpus of contemporary Catalan created as a continuation of the first COD compiled two decades before. In the first place, the new project intends to analyze the changes undergone by the Catalan language during this period. In second place, we no longer center the linguistic analysis on the traditional generativist techniques but on the tenets of Optimality Theory. And, finally, we use new dialectometric techniques based principally on probabilistic analytical procedures.


Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.

*Keywords:* Dialectology; Dialectometrics; Language variation; Linguistic distance; Optimality Theory; Oral corpus; Catalan

## 1. Introduction

In the mid 1990's the Department of Catalan Philology at the Universitat de Barcelona compiled and organized in databases an oral dialectal corpus (*Corpus Oral Dialectal*, i.e. COD) of contemporary Catalan (the whole description and results of the project are available at http://www.ub.edu/lincat). The corpus consisted of phonetic and morphological materials obtained through a 600-item questionnaire and a set of 10' spontaneous speeches. The survey was conducted in the 82 county seats –or equivalent towns– of the whole Catalan-speaking region with 3 middle-age informants at each location. The informants had not attended school in Catalan, belonged to the middle

* Corresponding author. Tel.: +34-93-5422410; fax: +34-93-5422321.
*E-mail address:* esteve.clua@upf.edu


Peer-review under responsibility of Universidad de Valladolid, Facultad de Comercio.
doi:10.1016/j.sbspro.2015.07.423

class, had no higher education, and had not spent lengthy periods away from their home. The criteria used for selecting the locations and speakers aimed at gathering information on the Catalan spoken by the majority of the population, who nowadays are mainly settled in urban areas (and not the Catalan spoken by older people in rural areas, as traditional geolinguistic studies do). The databases obtained from the questionnaire were published on a CD-ROM (see Viaplana, Lloret, Perea, and Clua, 2007 and also http://www.ub.edu/cccub/cod-cd-2007.html), and interactive maps based on these databases were made available at http://www.ub.edu/mapadialectal/, with the geographical distribution of the forms, and at http://www.iula.upf.edu/rec/vdm_val/, with dialectometric measures on Valencian varieties. Additionally, a selection of the spontaneous oral texts was published in the Digital Repository of the Universitat de Barcelona (http://diposit.ub.edu/dspace/handle/2445/10413), with the spoken audio files and their phono-orthographic and phonetic transcriptions.

Two of the main analytical goals of the COD project were, on the one hand, the study of linguistic variation in the areas of phonetics, phonology and inflection, and, on the other, the study and representation of the linguistic distance. The COD approach differs from other dialectometric analyses in the fact that these are just quantitatively surface-oriented, while we designed a new method (known as MCOD: *Mètode COD*) to capture the differences among varieties not only quantitatively but also qualitatively in order to increase the accuracy of the groupings (see, e.g., Viaplana, 1999; Clua, 1999; Clua and Lloret, 2006; Clua, Valls, and Viaplana, 2008; Clua, Lloret, and Valls, 2009; Valls, Nerbonne, Prokic, Wieling, Clua, and Lloret, 2012).

Since 2010, researchers from the Universitat de Barcelona, the Universitat Pompeu Fabra and the Universitat Autònoma de Barcelona have shared new projects on the study of the phonetic, phonological and morphological characteristics of Catalan (http://www.ub.edu/GEVAD), with the investigators of the two former groups specializing in gathering up-to-date data (i.e. COD2) and the dialectometric quantification of the data, using more recent techniques in both tasks. The questionnaire has been expanded to 700 items, but the interviews have been conducted in the same 82 towns with 3 informants with the same aforementioned characteristics (except for the fact that they have attended school in Catalan).

In the following sections, we will briefly present the novelties of our COD2 project with respect to the way we analyze the data (see § 2), how we introduce the temporal perspective in the spatial study (see § 3), and the new dialectometric techniques we use (see § 4).

## 2. Linguistic analysis

As mentioned, the MCOD incorporates a linguistic analysis of the data before applying the measure of distance for dialectometric studies. From this point of view, it is crucial to discriminate the unpredictable components of the language (i.e. the underlying phonological and morphological differences) from its predictable elements (i.e. the regular phonological phenomena that produce the phonetic outputs). In the COD studies, we analyzed these differences with traditional generative techniques, counting the underlying differences as well as the sum of phonological rules that give arise to different outputs. For example, as shown in (1), in Northwestern Catalan the 2nd person plural pronominal clitic has the underlying form /b+o+z/ in some varieties (1a), where /b/ is the '2nd person plural' clitic morph despite the plurality being transparent in the last morph (i.e. -/z/); but it has the underlying form /t+o+z/ in some others (1b), where /t/ is just the '2nd' (shared with the singular form) clitic morph. This underlying dissimilarity counts as 1 difference in our analysis.

| | | | |
|---|---|---|---|
| (1) | a. | /b+o+z/ '2nd person plural pronominal clitic'º | Cf. /t/ '2nd person singular pronominal clitic' |
| | b. | /t+o+z/ '2nd person plural pronominal clitic' | Cf. /t/ '2nd person singular pronominal clitic' |

See now the examples in (2), which illustrate surface proclitic differences from a unique underlying form of the 1st person singular pronominal clitic, /m/, in different Catalan varieties. The differences are due to the insertion of an epenthetic vowel due to syllabic reasons: the vowel is added either before or after the consonant and with the shape [ə], [e] or [a], according to the dialect. In classical generative phonology, each of these changes is formalized by

means of specific rules.

(2) /m/: [əm], [mə]; [em], [me]; [am], [ma] '1st singular pronominal clitic'

In (2) there is no difference in the underlying forms, but there is 1 difference for the pre- or post- position of the inserted vowel and 1 difference for the quality of the vowel. Hence, the pairs in (3a) show 1 difference, i.e. the difference regarding the quality of the preposed vowel; the pairs in (3b) show 1 difference, i.e. the difference regarding the quality of the postposed vowel, and the pairs in (3c) also show 1 difference, i.e. the difference regarding the position of the vowel. Instead, the pairs in (3d) show 2 differences, i.e. the difference in the quality of the vowel and in its position.

(3) a. [əm] – [em]; [əm] – [am]; [em] – [am]: 1 difference
b. [mə] – [me]; [mə] – [ma]; [me] – [ma]: 1 difference
c. [əm] – [mə]; [em] – [me]; [am] – [ma]: 1 difference
d. [əm] – [me]; [əm] – [me]; [əm] – [ma]: 2 differences

In COD2, the analysis we are pursuing is framed within the tenets of Optimality Theory (OT; Prince and Smolensky, 2004). With this new approach, there is no difference regarding the way we count differences due to the input (underlying) forms. However, there are significant differences in the way we count the differences due to surface phenomena, because in OT there are no rules to account for specific changes but rather the phonology of each language (or variety) derives from the way the universal constraints are ranked in each variety. Hence, surface differences are due to the differences among the constraint rankings that the varieties show (*factorial typology*). For example, in our example in (2), the Markedness constraints that govern vowel selection in epenthesis, which is a non-prominent (i.e. marginal: M) position, is the universal ranking presented in (4a), which disfavors more sonorants vowels in M positions: 'a' is worse (*less harmonic*, in OT terms) than 'e' and 'e' is worse than 'ə'. This ranking interacts with the universal ranking of constraints that, in general, favors more sonorous segments in the nucleus (N) position of the syllable (4b), because it is a prominent position: 'ə' is worse than 'e' and 'e' is worse than 'a'.

(4) a. *M/a >> *M/e >> *M/ə
b. *N/ə >> *N/e >> *N/a

In (5)-(6) we show how the OT analysis proceeds with some pairs. In order to capture the differences among the three varieties that prepose a vowel, we need the rankings in (5). With these orderings, now the differences between the pair [əm] – [am] amount to 12 (6a) and the ones between the pairs [əm] – [em] and [em] – [am] amount to 6 only (6b-c), whereas in our previous analysis all were dissimilar in 1 difference (cf. (3a)). Hence, our OT analysis can capture the dissimilarities among varieties in a more detailed scale and hopefully will allow for finer groupings in future work.

(5) a. [ə]-selection: *M/a >> *M/e >> *M/ə, *N/ə >> *N/e >> *N/a
b. [a]-selection: *N/ə >> *N/e >> *N/a, *M/a >> *M/e >> *M/ə
c. [e]-selection: *M/a, *N/ə >> *M/e, *N/e >> *M/ə, *N/a

(6) a. [əm] – [am]: 12 differences Cf. 1 difference in (3a)
b. [əm] – [em]: 6 differences Cf. 1 difference in (3a)
c. [em] – [am]: 6 differences Cf. 1 difference in (3a)

**3. Spatial and temporal variation**

Although only two decades have passed since collecting the linguistic data for the COD, we consider that due to different factors affecting the Catalan-speaking socio-linguistic community, this is a very important period of time

that requires in-depth study. Among other aspects, a number of factors have brought us to consider it important to create of a new corpus, i.e. the COD2. These include, on the one hand, the fact that in this period schooling in Catalan has become widespread in most of the territories within the linguistic area, above all in Catalonia and the Balearic Islands, but also in the Valencian Region; another issue is what has happened on the Aragonese Border, the territories in the French state where Catalan is spoken and the Sardinian city of Alguer, as in all these places the presence of Catalan in teaching has increased if we compare nowadays situation with the one experienced by our informants of the COD; and finally, due to the increased use of Catalan language in the written and audiovisual media, at least in some areas of the linguistic area.

Surely, these factors have favored an increase in de-dialectalization in many territories and the leveling tendency of the dialectal varieties. It is also possible that the so-called Border Effects have increased because the existence of different standards and their respective vertical convergence may produce horizontal divergence between them (see Valls, Wieling, and Nerbonne, 2013).

Thus this new corpus will serve, in the first place, to define the linguistic distance existing today between geographic varieties in the entire linguistic area, through a dialectometric analysis where we combine different quantitative analytical techniques. As can be seen in Fig. 1, this is an attempt to determine the spatial linguistic distances ($LD_e1$', $LD_e2$', $LD_e3$', $LD_e4$', etc) among the geographic varieties (V1', V2', V3', V4', etc) starting from the updated data in the COD2. And, later, through contrast with the dialectometric analyses of the COD (see Clua, Goebl, Casassas, Civit, and Salicrú, 2013), where the spatial linguistic distances were determined ($LD_e1$, $LD_e2$, $LD_e3$, $LD_e4$, etc) among the geographic varieties (V1, V2, V3, V4, etc) through the data previously collected with the corpus around 1994, we will be able to determine the temporal evolution of these linguistic distances ($LD_t1$, $LD_t2$, $LD_t3$, $LD_t4$, etc ) and, therefore, the linguistic change experienced during these years by the varieties of the Catalan language.

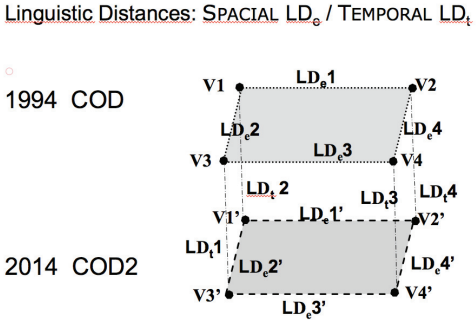


Fig. 1. Spatial/temporal linguistic distances

## 4. Dialectometric analysis

Regarding the dialectometric analysis which will be applied to the data from the COD2 in order to calculate the linguistic similarities and differences between the varieties of Catalan and thus to be able to determine the Linguistic Distance, we center principally on two methodologies: the MCOD and the Levenshtein Distance. As we have already mentioned, the MCOD is a method designed in the dialectometric analysis of the COD characterized principally by basing the quantitative analysis of data on a prior linguistic analysis. This methodology allows capturing the similarities and differences between dialectal varieties in a more suitable way. The algorithm used in this methodology is presented in (7).

$$(7)\ dist(i,j) = \frac{\sum_{k=1}^{long} dif_k(i,j)}{long} \times 100$$

Regarding the Levenshtein Distance, this is a measure to calculate the phonetic distance between two lines of data. To determine this distance, the Levenshtein algorithm searches for the smallest set of basic operations necessary to transform one line into another. These operations can be from inserting, removing or substituting segments.

We use the following tools to measure and graphically represent the Linguistic Distance: Visual Dialectometry (VDM) of the Dialectometric School of Salzburg, which allows for dendrometric and cartographic representation of the LD through different types of maps and algorithms (http://www.dialectometry.com/). Gabmap, an online application from the Center for Language and Cognition of the University of Groningen (CLCG) (http://www.gabmap.nl), one of the most powerful tools currently available for dialectometrization processes using the Levenshtein Distance. And finally, the DiaTech online application from the EUDIA Group of the University of the Basque Country (http://eudia.ehu.es/diatech/index/).

In contrast with the dialectometric analyses of the COD, based on deterministic methods, now we principally use probabilistic methods. The deterministic methods are oriented to classifying the populations (or varieties) in dialectal groups, which, with fuzzy or probabilistic methods, the classifications of the populations is complemented with the similarity of belonging to each group. Thus, while deterministic classification brings us to represent separate groups, fuzzy representation provides a more solid group structure and highlights the populations which live in the borders between the dialectal areas. This type of method allows us to better represent the linguistic reality of the populations who live in border regions between dialectal groups.

One of the fundamental and most novel objectives of the dialectometric analysis of the COD2 is centered on identifying the populations and linguistic forms which allow for the characterization of the dialectal varieties resulting from a classification process. Regarding the populations, the interest is on identifying those which constitute the reference population of each group (central or pattern populations). In the sphere of numerical classification, characterizing the groups is usually performed on the basis of considering these reference populations and the description of their differences.

With regards to the linguistic forms, we are interested in identifying the most informative forms both at the global level as well as for each of the groups. In this last case, we intend to determine the most representative and the most distinctive forms which allow us to characterize a dialectal group and differentiate it from the other groups. At the conceptual level, the information that a specific element or linguistic form (word, morpheme, phonic segment, etc.) provides can be associated to its discriminatory capacity for classification purposes. Ordering elements or linguistic forms by their informativity gives an idea of the importance of each of these in an individual manner. At the global level, its correlation must be taken into account in order not to center on a few aspects (which may be redundant) and to leave others aside, which though they may be less important, may also explain other dimensions of interest. For this purpose, ascending grouping algorithms are proposed for the forms by their information in order to find the most informative set of forms, and to highlight the structure of correlations between the elements.

## 5. Conclusion

In this article we have presented some of the principal and newest characteristics of the COD2, an oral dialectal corpus of contemporary Catalan which was created as a continuation of the COD with the primary goal of contributing to broaden the understanding of the geographic linguistic variation in general and on the distance between linguistic varieties, in particular. All this work is based on delimiting the linguistic distance between the varieties of the Catalan language but could be applied to other languages as well.

**Acknowledgements**

These results are part of the research projects FFI2013-46987-C3-3-P and FFI2013-46987-C3-1-P (sponsored by the Ministerio de Economía y Competitividad from the Spanish Government) and the research groups 2014SGR1317 and 2014SGR918 (sponsored by the Catalan Government). We thank the audience at CILC 2015 for their useful comments.

**References**

Clua, E. (1999). Distància lingüística i classificació de varietats dialectals. *Caplletra, 26*, 11 - 26.

Clua, E., and Lloret, M-R. (2006). New tendencies in geographical dialectology: The Catalan *Corpus Oral Dialectal* (COD). In J-P. Y. Montreuil (Ed.), *New perspectives on Romance linguistics, vol. 2 (Phonetics, phonology, and dialectology)* (pp. 31-47). Amsterdam/Philadelphia: John Benjamins.

Clua, E., Lloret, M-R., and Valls, E. (2009). Análisis lingüístico y dialectométrico del *Corpus Oral Dialectal* (COD). In P. Cantos Gómez and A. Sánchez Pérez, Aquilino (Eds.), *A survey on corpus-based research / Panorama de investigaciones basadas en corpus* (pp. 1033-1045). AELINCO. (CD-ROM.)

Clua, E., Valls, E., and Viaplana, J. (2008). Analisi dialettometrica del catalano partendo dai dati del COD. Una prima approssimazione alla gerarchia tra varietà. In G. Blaikner-Hohenwart, E. Bortolotti, R. Franceschini, E. Lörincz, L. Moroder, G. Videsott, and P. Videsott (Eds.), *Ladinometria. Miscellanea per Hans Goebl per il 65° compleanno. Edizione multiligue*, vol. 2 (pp. 27-42). Vigo di Fassa: Istituto Culturale Ladino.

Clua, E., Goebl, H., Casassas, X., Civit, S., and Salicrú, M. (2013). Anàlisi dialectomètrica del COD amb el suport del VDM. In E. Clua and M-R. Lloret (Eds.), *Qüestions de morfologia flexiva i lèxica del català* (pp. 133-168). Alacant: Institut Interuniversitari de Filologia Valenciana.

Prince, A., and Smolensky, P. (2004). *Optimality theory. Constraint interaction in generative grammar*. Oxford: Blackwell.

Valls, E., Nerbonne, J., Prokic, J., Wieling, M., Clua, E., and Lloret, M-R. (2012). Applying the Levenshtein distance to Catalan dialects: a brief comparison of two dialectometric approaches. *Verba, 39*, 35 - 61.

Valls, E.; Wieling, M., and Nerbonne, J. (2013). Linguistic advergence and divergence in north-western Catalan: a dialectometric investigation of dialect leveling and border effects. *Literary and Linguistic Computing, 28*, 1 - 28.

Viaplana, J. (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, J., Lloret, M-R., Perea, M-P., and Clua, E. (2007). *COD. Corpus Oral Dialectal*. Barcelona: PPU. (CD-ROM.)