DIALECTOMETRY: A NEW TREATMENT OF DIALECTAL MORPHOLOGICAL DATA

Maria-Pilar Perea

Universitat de Barcelona, Spain

1. INTRODUCTION

At the beginning of the twentieth century, Antoni M. Alcover (1862–1932), the founder of Catalan dialectology (Perea 2004a), embarked on a dialectal research project aiming to compile the conjugation of 75 regular and irregular verbs from 149 localities in the Catalan-speaking area. The study was published between 1929 and 1933, under the title of *La flexió verbal en els dialectes catalans*. The result is a complete, homogeneous, and systematic corpus of almost half a million verb forms. The data, collected between 1906 and 1928, reflect the status of the morphology of Catalan verbs at the beginning of the twentieth century.

The results of processing the information through a database which allowed rapid access to information were published in hardcopy and on CD-ROM (Perea 1999, 2004b). Later a computer program was designed in order to use these data to make computerized maps (Perea 2001). Both programs have been updated in 2005 (Perea 2005). The result of mapping is an atlas of the Catalan linguistic domain, which presents a thorough study of verb morphology. The object of this paper is to present a new treatment of Alcover's verbal data by using a quantitative approach. Applying the methodology of Hans Goebl (Goebl 1989, 1993), dialectometry gives the possibility of re-examining the specific realizations and considering the data from a general point of view. Here we limit our study to the forms of the first conjugation.

2. FROM LINGUISTIC GEOGRAPHY TO DIALECTOMETRY

The final aim of linguistic geography is to create a dialectal atlas. In each of the 6,000 maps in *La flexió verbal* (LFV) it is possible to draw dialectal borders that show the end and the beginning of the use of particular morphological forms, or areas where identical results overlap. The problem is that these results offer only one vision of the real situation. In linguistic geography, simultaneous study of the entire body of data is not possible. Dialectometry, however, analyses the linguistic reality from a global, generalizing perspective and avoids the problems posed by the idiosyncrasies of particular data (Goebl 2003:61).[1]

The first stage in the empirical and cartographic preparation of the LFV was to construct the Thiessen polygons, applying the principles of Delaunay-Voronoi geography. This is the only way to obtain an adequate drawing of combined isoglosses. LFV has 342 polygon edges, whose network is the graphic support of the dialectometric representations. Isoglosses are drawn with the help of a certain number of polygon edges (Goebl and Schiltz 1997).

Later, applying the dialectometric method to the data of LFV requires computerized treatment. The process is still in an initial, experimental phase, because we have used only, as a sample, the conjugation of the verb *cantar* 'to sing', which represents the first conjugation. In the future, we will leave out the additional verbs and new localities which appear in Alcover's notebooks but are not represented in their entirety. So, in a dialectometric treatment, if there are 149 localities, 149 answers are required. If the data are defective, the number of forms can never be lower than 30. As a whole, there will be 149 localities in the network — leaving aside the localities with very few data — and 75 verbs in total. In the future, a complete treatment of the whole set of data will use the results of 6,000 "working maps".

Before applying mathematical procedures to verbal forms, it is necessary to consider:

(a) Null responses. According to Goebl (1997: 23), missing data are always disturbing factors in taxometry. In LFV there are few cases of missing data. When they arise, they are attributed to ignorance of certain verbs (for example, the verb *caldre* 'to be necessary') that are not in general use throughout the Catalan domain. There are also cases of forgetting and mistakes.

(b) Multiple answers. Though they are not very frequent in Romance linguistic geography or in Romance linguistic atlases, Alcover's data show multiple answers in certain localities. Because multiple answers cannot be incorporated within the VDM program (Visual Dialectometry), it is necessary to choose the most representative form for each locality. With this constraint, and in order to obtain homogeneous data, variations have been eliminated.

(c) The pressure of standard Catalan. Goebl's dialectometric analysis of linguistic atlases habitually inserts an artificial atlas point of reference that represents the pressure on the dialect exerted by the standard variety. In the case of Alcover's data, this point is represented by the forms used in Barcelona; therefore, no new point is needed.

Next, the data in Alcover's original database (in Microsoft Access) were coded, a process carried out manually. This classification involves the grouping of results found in a dialect atlas map into types.

[1] I want to thank Hans Goebl and Slawomir Sobota for their generous help in the process of applying the dialectometric analysis to Alcover's data, and also in constructing an appropriate polygon network using computer-assisted cartography facilities.

The structure of the original database of LFV contained 14 fields. Thus it was necessary to give a number code to whole fields in Alcover's corpus. Without erasing any of the original information, the result had to be columns with a string of eight digits, which represented each "working map".

After this number coding, the second procedure was carried out at the University of Salzburg. Goebl used the distillation method: he constructed a smaller database, which contained a part of the aforementioned assigned figures. The resulting subdatabase was incorporated into the VDM program, a taxometric and cartographic program created by Edgar Haimerl.[2] The VDM program allows a range of calculations on the route from the data matrix to the similarity and distance matrix:

1. Measure of similarity between variables: There are several measures of similarity, each of which defines the similarity between places.
2. Numerical classification: These similarity matrices can be immediately visualized in VDM.
3. Similarity profile: A simple form, where one row of the similarity matrix contains the dialectal similarity of the reference location to all other localities. Neighbours with high similarity appear in red on the map, neighbours further away in orange, yellow, etc. and the furthest away of all in dark blue.

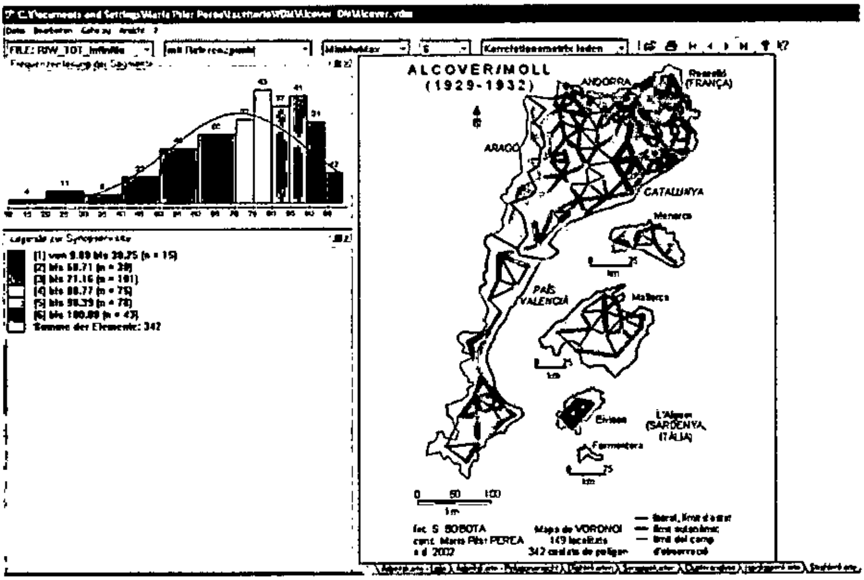
The VDM program allows immediate visualization of all the other similarity profiles stored in the respective similarity matrix, with a simple click. Thus, any locality may be selected and compared to each of the others. The similarity profile of another reference location can be displayed by clicking on another location in the interactive map. In this way, areas of dialect can be analyzed. One can find large areas with many immediate neighbours.

As well as this, synoptic evaluations of a similarity matrix are possible (for exanple, minimum, maximum, median, standard deviation, skewness). Different types of map (similarity profile, honeycomb maps, beam maps, cluster analysis dendrograms) can be presented. See, as a sample, Figures 1 and 2.

3. CONCLUSIONS

This presentation is a first attempt to apply a dialectometric analysis to a morphological Catalan atlas in a small sample of data, and for this reason it is still too early to obtain a reliable linguistic interpretation of the maps. However the future is very promising. Dialectometry will not only treat the data globally — something that is impossible with the individual representation of maps — but will also be able to determine and objectively classify the main Catalan dialects and subdialects from a morphologic point of view. However, there are still some questions to be considered: in Alcover's data, as was the case in the *Survey of English Dialects* (Goebl

[2] See `ald.sbg.ac.at/dm` for the program features and the software development.



FILE: RIW_TOT_Infinitiu

Korrelationsmatrix laden

Legende zur Synopsenkarte

[1] von 9.89 bis 38.75 (n = 15)
[2] bis 68.71 (n = 30)
[3] bis 71.16 (n = 101)
[4] bis 88.77 (n = 75)
[5] bis 98.39 (n = 78)
[6] bis 100.00 (n = 43)
Summe der Elemente: 342

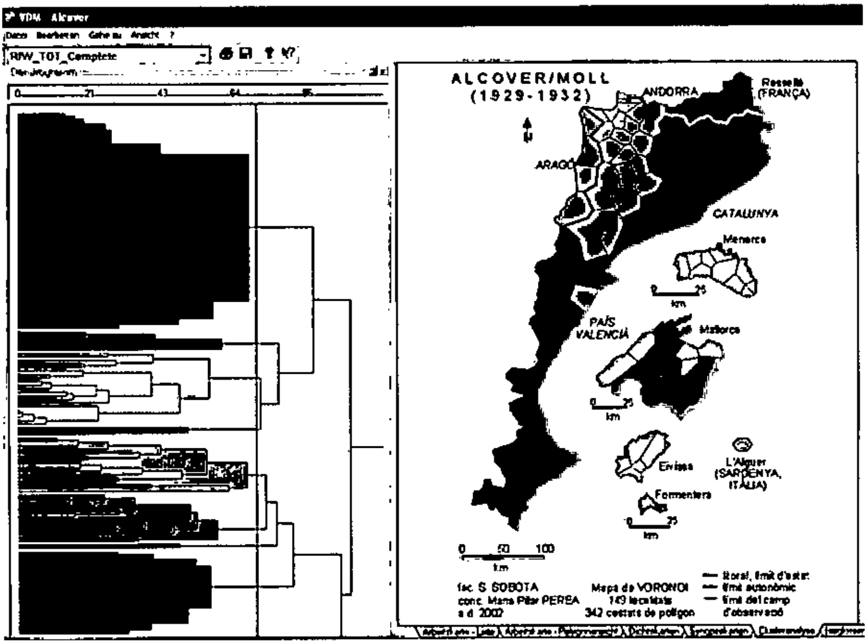
fac. S. SOBOTA
conc. Maria Pilar PEREA
a.d. 2002

Mapa de VORONOI
149 localitats
342 costats de polígon

— litoral, límit d'estat
— límit autonòmic
— límit del camp d'observació

FIGURE 1

"Beam map"



RIW_TOT_Complete

fac. S. SOBOTA
conc. Maria Pilar PEREA
a.d. 2002

Mapa de VORONOI
149 localitats
342 costats de polígon

— litoral, límit d'estat
— límit autonòmic
— límit del camp d'observació

FIGURE 2

Dendrogram

1997:23), there are localities with multiple answers. Choosing only one response is sometimes a difficult decision.

The application of mathematical and statistical methods to different sciences is common today. Linguistics has sometimes been criticized for the absence of application of a rigorous scientific method in its studies. However, the concept of science does not imply a sole category, which admits or rejects the inclusion of different areas of knowledge. In fact, any area of knowledge has to be judged on its own merits, evaluating its aims and the degree to which it is able to accomplish them. Since reality can be studied from different points of view, dialectometry offers a methodology and a number of techniques of numerical classification which are vital to our investigation of a new aspect of the complex dialectal reality.

REFERENCES

Alcover, A.M. and F. de B. Moll. 1929–1932. *La flexió verbal en els dialectes catalans*. Barcelona: Anuari de l'Oficina Romànica. V. II (1929) [73] 1–[184] 112, v. III (1930) [73] 1–[168] 96, v. IV (1931) [9] 1–[104] 96, v. V (1932) [9] 2–[72] 64).

Goebl, H. 1989. Problèmes et méthodes de la dialectométrie. In *New methods in dialectology*, ed. M.E.H. Schouten and P.Th. van Reenen, 165–184. Dordrecht: Foris.

——. 1993. Dialectometry: A short overview of the principles and practice of quantitative classification of linguistic atlas data. In *Contributions to quantitative linguistics*, ed. R. Köhler and B.B. Rieger, 277–315. Dordrecht: Kluwer.

——. 1997. Some dendrographic classifications of the data of CLAE 1 and CLAE 2. In Viereck and Ramisch, pp. 23–32.

——. 2003. Regards dialectométriques sur les données de l'Atlas Linguistique de la France (ALF) : Relations quantitatives et structures de profondeur. *Estudis Romànics* XXV:61–117.

Goebl, H. and G. Schiltz. 1997. A dialectometrical compilation of CLAE 1 and CLAE 2: Isoglosses and dialect integration. In Viereck and Ramisch, pp. 13–21.

Perea, M.-P. 1999. *Compleció i ordenació de* La flexió verbal en els dialectes catalans (*A.M. Alcover i F. de B. Moll*). Barcelona: Institut d'Estudis Catalans (+ CD-ROM).

Perea, M.-P. 2001. *La flexió verbal en els dialectes catalans d'A.M. Alcover i F. de B. Moll. Les dades i els mapes*. Palma de Mallorca: Conselleria d'Educació i Cultura, Govern de les Illes Balears (CD-ROM).

Perea, M.-P. 2004a. The history of a multidialectal Catalan dictionary: "The Diccionari català-valencià-balear". In *Historical dictionaires and historical dictio-*

nary research, ed. J. Coleman and A. McDermott, 109–118. Tübingen: Max Niemeyer.

Perea, M.-P. 2004b. New techniques and old corpora: 'La flexió verbal en els dialectes catalans' (Alcover-Moll, 1929–1932). Systematisation and mapping of a morphological corpus. *Dialectologia et geolinguistica* 12:25–45.

Perea, M.-P. 2005. *Dades dialectals. Antoni M. Alcover.* Palma de Mallorca: Conselleria d'Educació i Cultura, Govern de les Illes Balears (CD-ROM).

Viereck, W. and H. Ramisch. 1997. *The computer developed Linguistic Atlas of England* 2. Tübingen: Max Niemeyer.