CATALAN GEOLINGUISTICS AND NEW TECHNICAL PROCEDURES

Maria-Pilar PEREA

Universitat de Barcelona

mpilar.perea@ub.edu

Abstract

New technologies are helping researchers to apply new methods to the treatment of dialectal data, accomplishing a variety of research objectives in the stages of data compilation, data processing and the presentation of results. In this regard, dialectology has at least two aspects: *a*) obtaining new data to learn contemporary linguistic variation; and *b*) retrieving earlier material to facilitate the study of linguistic change and create new opportunities to compare such data with present-day results.

The main aim of this paper is to show different techniques applied to various sorts of earlier dialectal materials in order to improve the cartographical representation of the data. For the Catalan dialects, a process of mapping has been applied to data gathered by Antoni M. Alcover roughly a century ago in *La flexió verbal en els dialectes catalans* ("Verbal flexion"), which is a complete corpus of the verb morphology of Catalan. We present new resources applied to this dialectal data, using dialectometrical calculations with nearly half a million verbal forms and retrieving the vocal sound of morphological forms through voice synthesis.

In addition, we show a new pattern of dynamic cartography that combines early and modern lexical data. Referring not only to space but also to time we use two axes of linguistic representation and, because of linguistic change, the representation of the words in space is layered. For that reason, we can speak of "dialectal stratigraphy".

An important advantage of the three procedures described in this paper is that the methodology can be adapted to other sorts of dialectal data for other languages.

Keywords

Dialectology, geolinguistics, Catalan, automatic mapping.

1. Introduction

Today new technologies are helping researchers to apply new methods to the treatment of dialectal data. These technologies are useful for accomplishing various objectives. For instance, in the data compilation stage, notebooks have been replaced by audio-taped or video-recorded interviews, which can be stored and processed easily. Moreover, a set of enquiries now makes up an "archive" of sounds or images that can be used to obtain a variety of results.

In the data processing stage, standard databases provide a useful tool to systematise and order dialectal materials and they can help in the presentation of results as well, whether in paper format or through automated cartography. Automated mapping is a desirable method as well for presenting large amounts of data. It offers an interactive tool that makes the representation of data easier in a given space.

Dialectology has at least two aspects: *a*) gathering new data to identify contemporary linguistic variation; and *b*) retrieving earlier material to facilitate the study of linguistic change and create new opportunities for comparing such data with present-day results. From that perspective, the main aim of this paper is to show the application of different techniques to various sorts of dialectal materials in order to improve the cartographical representation of the data. For the Catalan dialects, mapping has been applied to two types of materials: *a*) recent materials taken from surveys made by the ECOD project (Clua, Lloret & Perea 2006); and *b*) earlier materials gathered by Antoni M. Alcover approximately a century ago in *La flexió verbal en els dialectes catalans* ("Verbal flexion"), a complete corpus of the verbal morphology of Catalan. We present new resources applied to this earlier dialectal data: (1) using dialectometrical calculations with nearly half a million verbal forms; and (2) retrieving the vocal sound of morphological forms through voice synthesis. Moreover, we show a new pattern of dynamic cartography that combines early and modern lexical data. Making reference not only to space but also to time, we use two axes of linguistic representation and, because of linguistic change, the representation of words in space is layered. For that reason we can speak of "dialectal stratigraphy".

2. Reviewing earlier dialectal materials

Catalan did not have its first dialectal atlas until 1964 with the *Atlas Lingüístic de Catalunya*, by Antoni Griera. Sixty years earlier, Antoni M. Alcover, the father of Catalan dialectology, started linguistic surveys in order to compile the lexical, phonetic, morphological and syntactic idiosyncrasies of the different territories. However, his main aim was not to make a dialectal atlas, but rather a dictionary (the *Diccionari català-valencià-balear* (DCVB) (1930-1962)), which was completed with the collaboration of Francesc de B. Moll (Perea 2004). As a result, the surveys that gathered lexical information were not designed to be mapped, but rather to build an enormous lexicographic work in ten volumes, which can now be consulted on the internet: <http://dcvb.iecat.net/>.

From the Alcover surveys, two types of data have been recovered: dialectal information from 65 notebooks (the so-called *Quaderns de camp* (Perea 2006) and the verbal forms in the work *La flexió verbal en els dialectes catalans* (Perea 2005), which we will look at in more detail. In the latter work, the systematic presentation of verbal morphology has made it possible to: *a*) apply a quantitative approach in order to analyse the data dialectrometrically; and *b*) reconstruct the sounds of the phonetic transcriptions.

2.1. Making use of Catalan morphological verbal data: La flexió verbal en els dialectes catalans

Alcover began his dialectological survey in 1906. His methodology is described in several documents. In the early stages of the enquiry he applied the traditional methods of dialectology, looking for old informants (without making distinctions between men and women), who could produce the most genuine forms of the language. The questionnaire was extremely difficult — subjects had to give the complete conjugation of more than 77 verbs — and the dialectologist was obliged to change his mind and study young subjects ranging in age between 10 and 14. Alcover justified this change by saying that children and adolescents were free of linguistic prejudices. The subjects were questioned in groups of four or six.

In 1928, the year of the last enquiry, his collaborator, Francesc de B. Moll, started the work of ordering, grouping and processing all the data included in the field-books.

He modified the presentation of the information to adapt it to be published. The aim was to synthesise this vast body information giving each locality a number, and recording the orthographic and phonetic representation of the endings for each person of each verb. The reference form is the infinitive, but some second conjugation verbs take the reference form the root of the verb.

The ordering carried out by Moll was extremely valuable. It offered researchers the opportunity to consult a very big set of data (four hundred thousand verb forms) in only 368 pages. But it is also true that the data are not easy to search, and in fact the excessive simplification prevents readers from exploiting all its possibilities. The new computerised presentation overcomes these practical difficulties.

Although "Verbal flexion" was not originally conceived as a linguistic atlas — probably because of the logistical difficulty presented by such a large amount of data — the materials of this morphological study form a corpus for creating a computerised linguistic atlas. In 2000, a project, sponsored by the Balearic Government, used this material to develop a computer mapping program. The CD-ROM *La flexió verbal en els dialectes catalans d'A. M. Alcover i F. de B. Moll. Les dades i els mapes*, published in 2001, includes two computer programs: "*La flexió verbal en els dialectes catalans*" (the database) and "Les dades i els mapes de *La flexió verbal*" (the program that design maps on the screen). In 2005 a new CD-ROM version was edited to avoid incompatibilities and problems of visualisation arising from the ongoing updates of Windows versions (97, 2000, Millennium, XP).

In fact, the only way to treat the 470,255 verb forms was by applying computerised technology. The number of possible general maps, if we multiplied the total number of verb forms (55) that each verb contains and the total number of verbs studied (117), surpassed 6,000. This is the number of pages that a linguistic atlas of the corpus would need. Making a conventional dialectal atlas is impossible today, because of its dimensions, its cost, and the time the task would require. Using the computational program the user designs his own set of maps — general, or according to dialectal area — to make up a morphological atlas.

In addition to automated mapping, "Verbal flexion" can be exploited by using two new procedures: dialectometry and voice synthesis.

2.1.1. Dialectometry

The final aim of linguistic geography is to create a dialectal atlas. The unit of dialectal classification on a map is the concept of isogloss. In each of the 6,000 maps in *La flexió verbal* 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, generalising perspective and avoids the problems posed by the idiosyncrasies of particular data (Goebl 2003: 61). This methodology also highlights internal grouping and structures from the linguistic data that direct observation does not identify; through quantitative methods, it uncovers the profound structures that can be extracted from the superficial structures (Goebl 2003: 61).

The complete, systematic morphological corpus of *La flexió verbal* facilitates the application of dialectometric analysis. This method will produce interpretative maps by measuring, summarising, and simplifying the data. Dialectometric analysis is also useful to classify dialectal areas, by replacing the concept of isogloss as a basic unit of dialectal classification with the concept of linguistic distance. The isogloss is understood as an ideal line that signals, on a linguistic map, the limit between the presence or the absence of a given feature. The linguistic distance is related to the quantification of similarities between the linguistic realisations of two local dialects.

The first stage in the empirical and cartographic preparation of "Verbal flexion" was to construct the Thiessen polygons, applying the principles of Delaunay-Voronoi geography. Later, using the dialectometric method to the data of "Verbal flexion" requires computerised treatment. Next, the procedure of adapting the materials of Alcover's original database (in Microsoft Access) starts, in order to taxate or code the data.

After this number coding, the next procedure was carried out at the University of Salzburg. Dr. Goebl used the distillation method; that is to say, he constructed a smaller database, which contained a part of the aforementioned assigned figures. The subdatabase resulting was incorporated to the VDM program. The VDM, a taxometric and cartographic program created by Edgar Haimerl, is a tool kit that provides users

with a range of methods and algorithms and enables them to analyse highly specific aspects of their data (see <http://ald.sgb.ac.at/dm> for the program features and the software development).

The VDM program allows a range of calculations on the route from the data matrix to the similarity and distance matrix. It allows also immediate visualisation of all the other similarity profiles stored in the respective similarity matrix, with a simple click. Synoptic evaluations of a similarity matrix are possible (e.g. minimum, maximum, median, standard deviation, skewness) and different types of map (similarity profile, honeycomb maps, beam maps, cluster analysis dendrograms) can be presented (see figures 1 and 2).
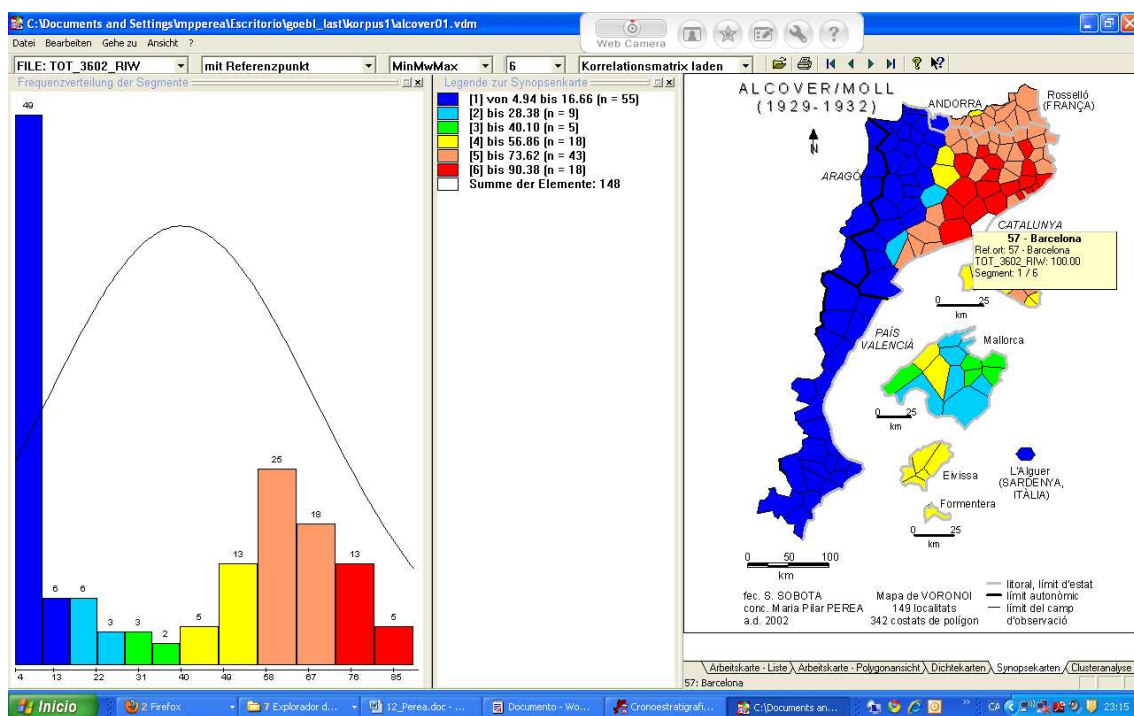
Figure 1. A similarity profile of the Catalan linguistic domain: similarity map to the "Verbal flexion" locality Barcelona

152
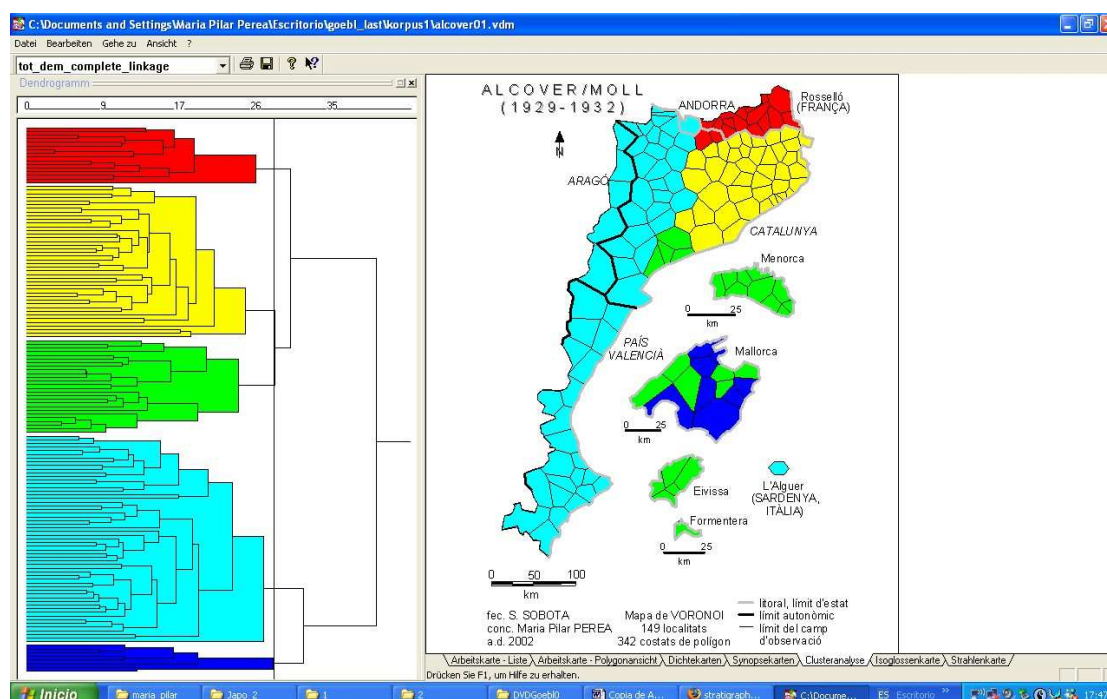
Figure 2. A dendographic representation of the total data set of the "Verbal flexion"

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. Furthermore, cluster analysis, with its diachronic interpretation, may provide evidence for the existence of a unique, undiversified linguistic stage, or for the hypothesis supported by some scholars that Catalan was born from a dialectal double differentiation (Eastern and Western).

2.1.2. Voice synthesis

To get maximum efficiency out of the materials in "Verbal flexion", the Balearic Government has funded experimentation using voice synthesis techniques. Since the corpus in question incorporates a phonetic transcription of each register, it is not impossible to develop a program allowing voice synthesis, i.e. the automated reading aloud of a sound sequence corresponding to a register (or to a set of identical registers), using the database of "Verbal flexion". The reading requires transforming a sequence of

phonetic signs — gathered from different informants' pronunciation of about 4,000 syllables — into an acoustic sign understandable to a speaker of the language. To reach this objective, the phonetic signs have to be associated with the corresponding sound in each verbal form, and the final output is the generation of a synthesised voice, which can be, depending on the user's wishes, either masculine or feminine (Perea 2010).

To present the results, the synthesised sequence is placed alongside the phonetic realisations that appear in each of the maps generated automatically. Now the results can consulted on the internet: "*La flexió verbal en els dialectes catalans* d'Antoni M. Alcover i Francesc de B. Moll: Les dades, els mapes i la veu" <http://www.grubit.net/sintesi/> (see figures 3 and 4). When selecting the results, as a function of a given search, each register can be activated to obtain the corresponding sound sequence. Thus, different sound maps produced by synthesised voice are generated.

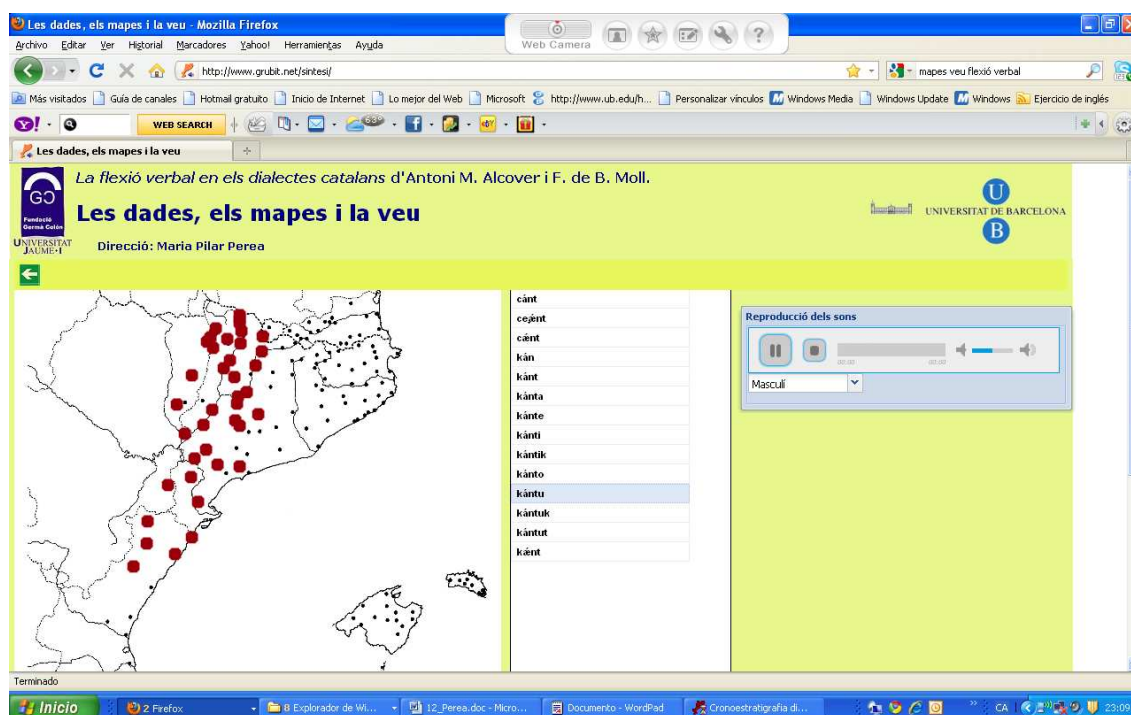Figure 3. Applying voice synthesis to "Verbal-flexion" <http://www.grubit.net/sintesi/>

154

Figure 4. A synthetic sound-geographic representation of the 1st person indicative
of *cantar* 'to sing'

Interest in the project is fourfold. Firstly, it involves using voice synthesis to recreate or reconstruct the locutions of informants who were interviewed more than a century ago. It is necessary to note that Catalan does not have sound registrations from dialectal surveys until the second half of the twentieth century. Secondly, the materials, despite their age, make comparative studies of diachronic typology possible, and the new application facilitates interactivity between the user and the program. Moreover, from an educational perspective, the project promotes knowledge of and respect for dialectal varieties that are remote from the standard language and it makes it possible to analyse verbal forms that have now disappeared. Lastly, it involves the creation of improved tools to treat dialectal data of diverse characteristics, whose results can be extrapolated to other computer applications that require similar technologies.

2.2. Dynamic cartography with diachronic data: Dialectal stratigraphy

In the Catalan dialects, as in other languages, lexical "diatopic" variations coexist. Different words may designate an identical semantic reality, despite having different etymologies (for example, "arena" vs. "sorra" [sand], "anyell" vs. "corder" [sheep] or "enemic" vs. "padrastre" [hangnail]). Moreover, they can extend not only over different geographical areas but may appear in different time periods.

Diatopic variation does not only affect two words. Sometimes four or more words can live together. For instance, "ham" can be designated in Catalan by at least four words — *cuixot*, *pernil*, *bacó* and *perna* — which are used depending on the geographical area and the time period involved.

Today a Catalan speaker knows that the word "atlot" [boy] is used in Majorca with the same meaning as the Valencian word "xiquet" or the Catalan word "nen". Capturing these data synchronically in a map does not present too great a problem. However, what happens to data of a historical character? Looking at Catalan from an etymological point of view — provisionally, due to the lack of access to complete literary and non-literary documentation from the origins of the language — it is possible to determine the first chronological use of certain words and their geographical location. Thanks to Colón's studies (1989: 283-296), we know, for instance, that the word "enemic" designated from the sixteenth century a small, bothersome piece of skin that is formed near a fingernail (a hangnail) not only in the area of Catalonia but also in Aragon and Navarra. In the Balearic Islands and in Valencia, by contrast, this bothersome piece of skin was designated "padrastre" from the nineteenth century, as in Portugal and Castile. In Catalan "enemic" and "padrastre" have been substituted by "reveixí" (19th century), "repèl" (19th century) or "repeló" (20th century) and sometimes by the learned word "cutícula". How can this variation in space and time be reflected cartographically?

The project of "dialectal stratigraphy", carried out in collaboration with Germán Colón under the Balearic Government's auspices, focuses on cartography of the historical lexicon. It seeks to represent cartographically the evolution which a sample of more than fifty words showing diatopic variation has experienced, simultaneously showing space and time axes. The project starts from the earliest period of written

156

documentation and shows the words' geographical location and their formal and semantic progression over time (Perea & Colón 2010).

From a methodological point of view, a number of dynamic maps have been created. The maps have a chronological axis that: a) permits location of the word (or words), starting from its origins; b) shows its evolution over the passage of the centuries; and c) makes it possible to observe its geographical distribution.
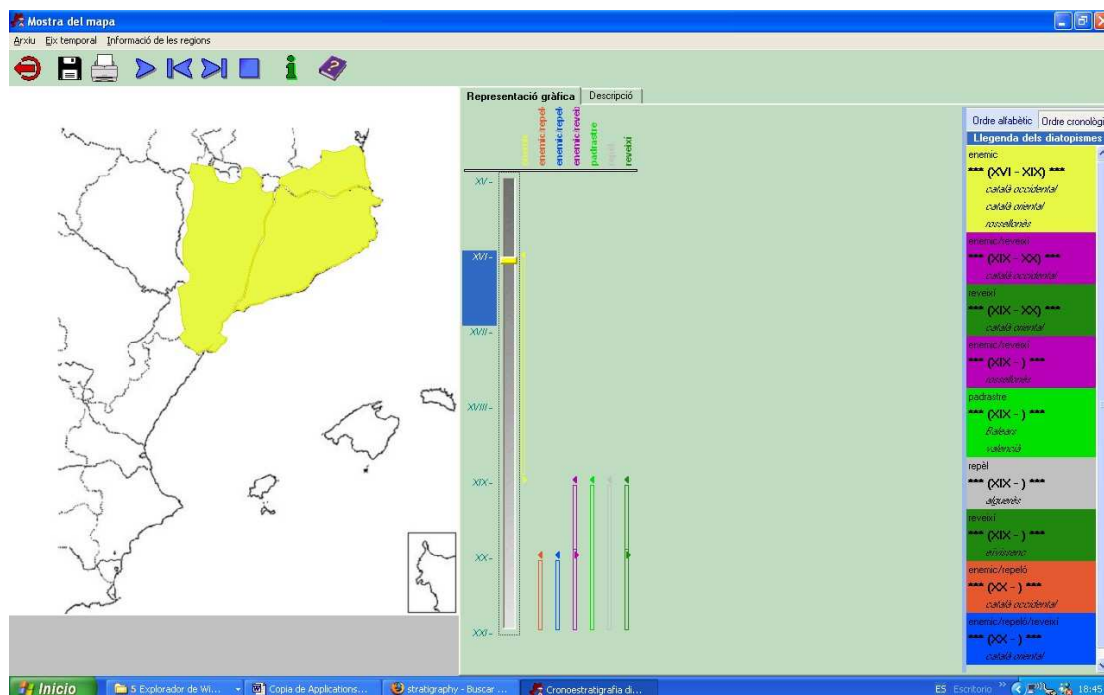

Figure 5. Dialectal stratigraphy: The Catalan word "repeló" (*hangnail*) in the 16th century

In geology, the concept of stratigraphy involves layers that are superimposed. Such superimposition may be possible in dialectal stratigraphy, but it does not always happen. It is not as straightforward in this case to proceed to the substitution of a layer of one colour by another layer of a different colour.
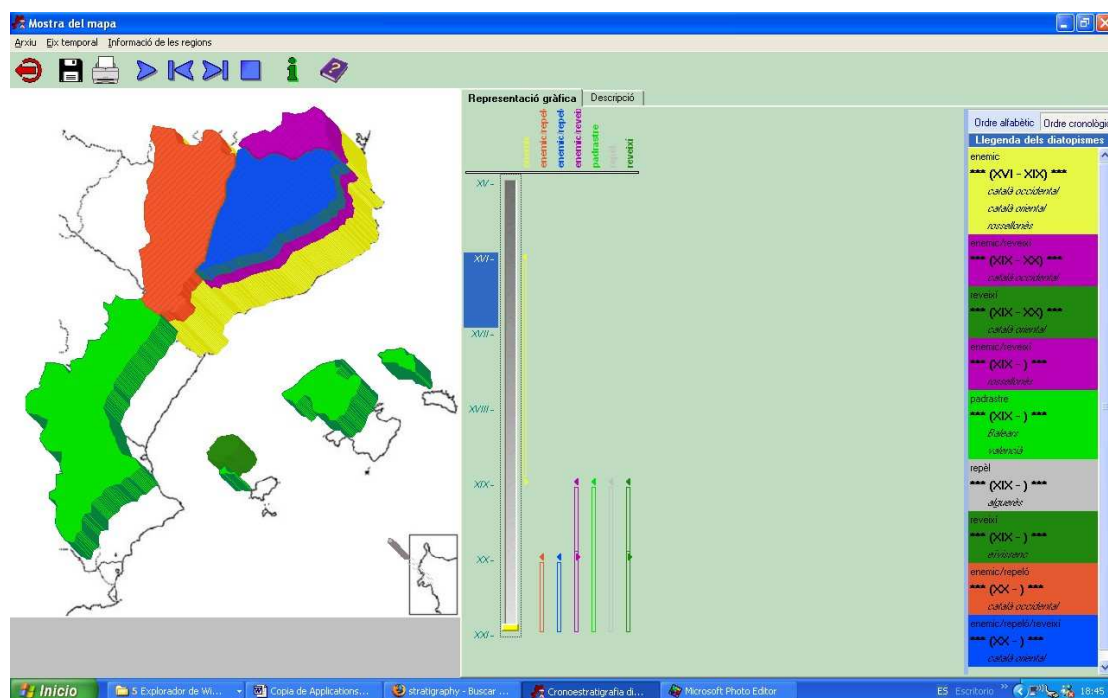
Figure 6. Dialectal stratigraphy: The chronological evolution of "repeló" (*hangnail*).

Sometimes a word does not completely replace another, but rather both of them coexist in the same area. In that case, we have opted for the use of a layer whose colour is the result of the mixture of two different colours. The layer then shows a duality of answers in the same geographical area [cucala vs. cornella 'crow' in Valencia and the Roussillon before the 14th century]. Since the data do not come from surveys made at specific geographical points, the dialectal areas are associated with Catalan's six main dialects in general terms (Rossellonès, Eastern Catalan, Western Catalan, Valencian, Balearic and Alguerès). Sometimes, depending on the distribution of the words, smaller geographical areas have been determined or words have been restricted to certain towns (for instance, the Vall d'Albaida in Valencia, or the eastern fringe of Aragon).
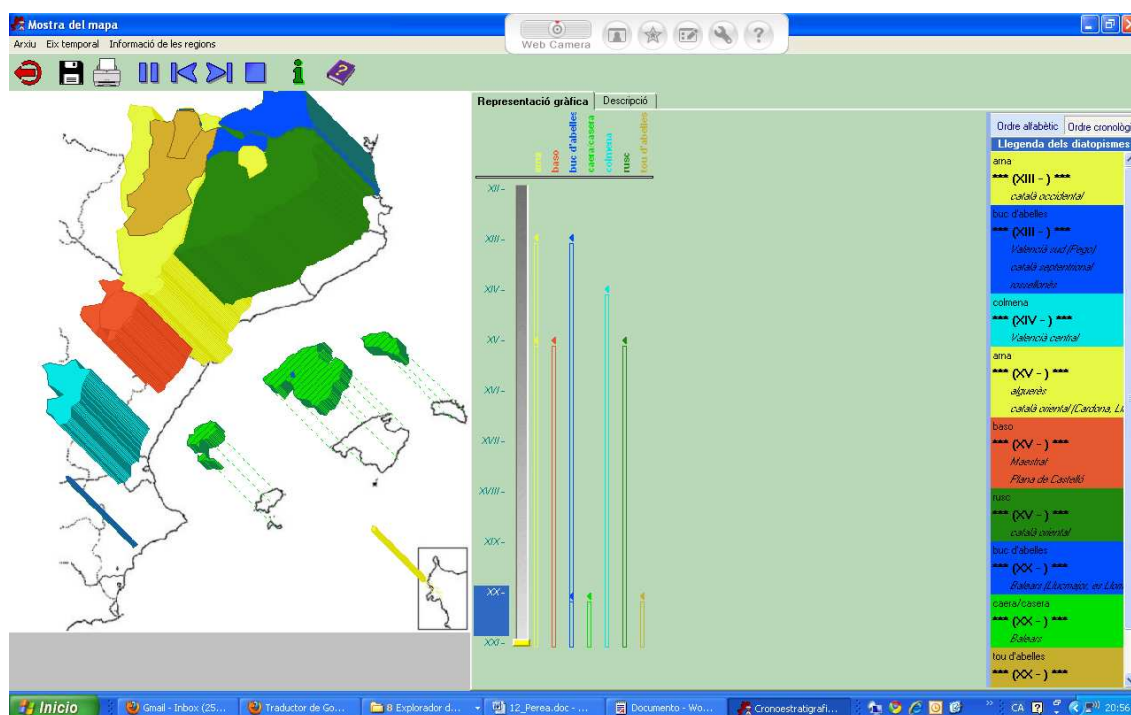
Figure 7. Small geographical areas representing the diatopic evolution of the word *arna* (*buc d'abelles*, *rusc*, *colmena*, *baso*, *casera*) 'honeycomb'

3. Conclusion

Voice synthesis, dialectometry and dialectal stratigraphy are only three examples of the new methods and technologies being applied to linguistic data. Although in this case they have been used with early data, the three procedures can be adapted to other sorts of data. Voice synthesis, however, is only useful when phonetic transcription is available. With present-day recordings a new technique can also be developed: voice recognition. Voice recognition will enable easier phonetic and orthographic transcription of oral recordings, and perhaps in the not-too-distant future, we will be able to develop the procedure and present the tool.

References

ALCOVER, Antoni M. & Francesc de B. Moll (1929-1933) *Diccionari català-valemcià-balear*, Palma de Mallorca: Moll (electronic edition <http://dcvb.iecat.net/>).

CLUa, Esteve, Maria-Rosa LLORET & Maria-Pilar PEREA (2006) "How to exploit a corpus: The experience of the Catalan 'Corpus Oral Dialectal", in *Proceedings of the 4th International Congress of Dialectologists and Geolinguists*, Riga: Latvian Language Institute, 102-111.

COLÓN, Germán. 1989. *El español y el catalán, juntos y en contraste*. Barcelona: Ariel.

GOEBL, Hans (2003) "Regards dialectométriques sur les donnés de l'Atlas Linguistique de la France (ALF): Relations quantitatives et structures de profondeur", *Estudis Romànics*, XXV, 61-117.

GRIERA, Antoni (1923-1964) *Atlas Lingüístic de Catalunya*, Barcelona: Institut d'Estudis Catalans/Ediciones Polígrafa.

PEREA, Maria-Pilar (2004) "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, Maria-Pilar (2005) *Dades dialectals. Antoni M. Alcover*, Palma de Mallorca: Conselleria d'Educació i Cultura. Govern de les Illes Balears, (CD-ROM edition).

PEREA, Maria-Pilar (2006) "From the notebook to the computer: the systematisation of a dialectal corpus", in *Proceedings of the 4th International Congress of Dialectologists and Geolinguists*, Riga: Latvian Language Institute, 400-412.

PEREA, Maria-Pilar (2006) "Dialectometry: A new interpretation of dialectal morphological data", *Linguistica Atlantica*, 27-28, 86-91.

PEREA, Maria-Pilar (2010) "Retrieving the sound: applying speech synthesis to dialectal data", in xxxxx, *Proceedings of Methods XIII. Papers from the Thirteen International Conference on Methods in Dialectology*, *2008*, Barry HESELWOOD, Clive UPTON (ed), Frankfurt: Peter Lang, 143-152.

PEREA, Maria-Pilar & Germán COLÓN (2010) "Cronoestratigrafía dialectal", in Maria ILIESCU, Heidi SILLER-RUNGGALDIER, Paul DANLER (ed), A*ctes du XXVe Congrès International de Linguistique et de Philologie Romanes,* Berlin: Mouton de Gruyter, IV, 199-211.