Dialectologia 2 (2009), 23-35.

NEW INSIGHTS INTO THE USE OF VDM: SOME PRELIMINARY STAGES AND A REVISITED CASE OF DIALECTOMETRY

Marcela J. Rivadeneira[1] and Xavier Casassas[2]

[1] Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra

marcela.rivadeneira@upf.edu

[2] Fachbereich Romanistik, Universität Salzburg

xavier.casassas@sbg.ac.at

Abstract

This paper intends to provide some basic guidelines for the preparatory stages before using 'Visual Dialectometry' (VDM). A detailed summary of the main steps that should be followed in order to create a VDM project is presented along with a revised case of Dialectometry using Viaplana's work *Entre la dialectologia y la lingüística,* whose results have been reconsidered under different algorithms. The dialectometrization of Viaplana's data has been treated with the methodology of the Salzburg School. The use of the VDM software has shown similar results to those in Viaplana's work when applying UPGMA, Complete Linkage, and Ward algorithms in a cluster analysis. However, further philological research is necessary in order to explain the differences in the resulting dialect groupings.

Keywords

Dialectology, dialectometry, VDM, cluster analysis.

1. Introduction

This paper focuses on the stage of data preparation to be employed in Visual Dialectometry (VDM), and intends to be a summarized guidelines for those researchers who might want to develop a DM project using Salzburg's methodology. Having considered this, we have taken Viaplana's work *Entre la dialectologia y la lingüística* as an example of data treatment. In his research, Viaplana et al. (1999) compiled dialectological material regarding the Northwestern varieties of Catalan. The branches analysed included Phonetics, Phonology, and Morphology, and data were collected through the use of linguistic interviews. The main purpose of his work was to determine

the state of the language in the major Northwestern urban areas – and expectedly more evolved varieties – and to compare them with previous states of the language in order to establish an evolution line for those varieties in the last decades.

The multiple features of VDM have been under constant development since 2000, when it was first conceived[1]. VDM is intended to be a primary tool for the measure and visualization of linguistic similarity and distance between dialects in the field of Dialectometry (DM). As regarded by Goebl (2006: 423): 'The Salzburg version of DM represents a heuristic instrument of explorative data analysis of universal applicability, on account of ist manyfold instruments of analysis developed on the job (ranging from the similarity maps to correlative DM)'.

VDM employs Salzburg's DM methodoloy in taxometric and cartographic analyses for the exploitation of linguistic atlases and geolinguistic research. VDM's objective (Goebl 2006) is both numerical – with the use of various statistics computations, and graphical – in order to visualize numerical results employing different colours.

There are two possible ways of estructuring the input of data in order to work with VDM:

(1) Database: It is necessary to account for an appropriate database consisting of a specific structure with the following tables: Sites (sampled/inquiry points), Title (N° map/map attributes = working maps), Taxates, and Data (linking the answer with the taxate and the map). All this information is treated in an Access database. The content of the tables and the relational diagram can be seen in Figure 1 and an example of a Sites table is shown in Figure 2.

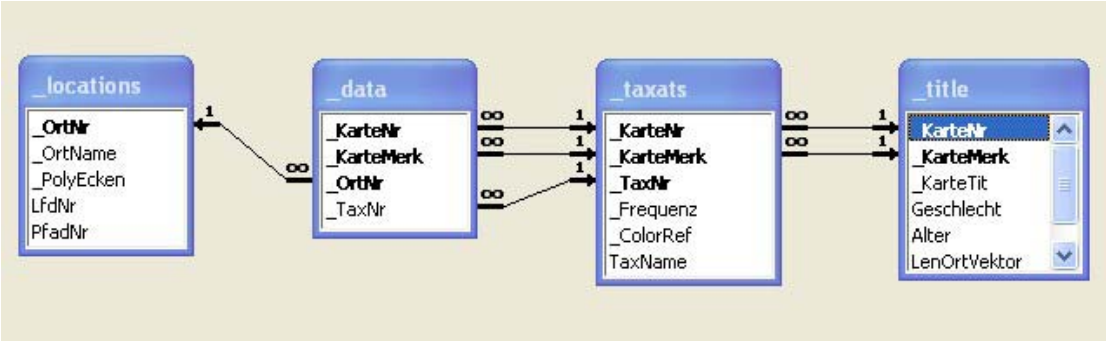

Figure 1. Main fields included in the database. English translation for the boldface fields in order of appearance: No. site, No. map, map attribute, No. taxate.

[1] The first version of VDM was developed by Edgar Haimerl, Hans Goebl's senior research assistant.

	_OrtNr	_OrtName	_PolyEcken
+	-1	bounding box	0,0,15360,11136
+	1	La Seu d'Urgell	8040,3902;8247,2189;8260,2210;8325,2247;8
+	2	Sort	7221,1903;7269,1900;7368,1950;7435,1942;7
+	3	El Pont de Suert	5757,3270;5914,2253;6084,2199;6224,2214;8
+	4	Tremp	6517,3948;7232,3365;7869,4176;7840,4576;7
+	5	Balaguer	6400,5189;6424,4995;6672,4502;7509,4810;7
+	6	Lleida	6020,5623;6400,5189;6827,5742;6542,6694;8
+	7	Les Borges Blar	6542,6694;6827,5742;7263,5582;7745,6904;8
+	8	Tàrrega	7263,5582;7509,4810;7776,4641;8000,6671;7
+	9	Cervera	7776,4641;7840,4576;8319,4915;8000,6671;7
+	10	Fraga	5169,7056;5418,5454;6020,5623;6297,6717;8
+	11	Benavarri	5561,4533;5757,3270;6517,3948;6672,4502;8
+	12	Tamarit de Llitera	5418,5454;5561,4533;6424,4995;6400,5189;8
+	13	Vall-de-Roures	5609,9116;5191,8951;5102,8937;5016,8904;8
+	14	Gandesa	5200,7069;6150,6846;6271,8005;5724,8204;8
+	15	Móra d'Ebre	6150,6846;6297,6717;6519,6707;6925,8344;8
+	16	Falset	6519,6707;6542,6694;7745,6904;7929,7058;7
+	17	Tortosa	7202,8832;7107,8919;7033,8967;6919,9008;8
+	18	Estàndard	7840,4576;7869,4176;8040,3902;8538,3795;8
▶	0		

Figure 2. Database representation regarding *Sites* and their specific geographic coordenates.

The database also includes other tables which are necessary for a proper internal processing in VDM and for filing an important part of the data obtained throughout the analyses, for instance, information on the different clustering methods employed.

At Salzburg laboratory, other projects are also processed using external databases from other researchers. Nevertheless, as the data structure may probably differ from the one employed here, it is necessary to convert the data into the system requirements; previous analyses concerning the structure and characteristics of the information are necessary in order to develop an adequate software that can convert the data correctly. One such a case is the database on Alcover and Moll's *La flexió verbal en els dialectes*

catalans (1929-1932) automated by Perea (2001) and adapted for the structure of VDM[2].

Every invited researcher may use the working methodology system that s/he considers the most apropriate. However, her/his data shall be adapted into the structure presented here.

Further methods that could be employed with Salzburg DM for the filing and processing of databases will not be mentioned here for reasons of space.

(2) Symmetrical similarity matrixes: They are employed at a first stage when no database is available. To begin, the similarities have to be measured, and 'As the dialectometrician can use different similarity measures, s/he has to select the appropriate index, keeping in mind her/his theoretical hypotheses on dialectal similarity' (Goebl 2007: 137). Complete matrixes are required to start processing a new VDM project (see Figure 3), and in this case it is also possible to use distance matrixes from other researchers who have employed different dialectometric methods and mathematical algorithms – such as the Levenshtein, successfully used by Heeringa with data from the *Reeks Nederlandse Dialectatlassen* in a VDM project during a research stay in 2007 at Salzburg laboratory. Once these matrixes are adequately adapted and converted (see below the methodology employed in this study) they can be analysed in VDM.

ROWTYPE_	VARNAME_	A33	A23	A50	A1	A38	A3	A18	A52	A19
PROX	A33	100.00	93.30	90.57	90.51	91.99	90.33			
PROX	A23	93.30	100.00	93.54	93.76	95.07	93.85			
PROX	A50	90.57	93.54	100.00	95.53	96.95	96.48			
PROX	A1	90.51	93.76	95.53	100.00	96.24	95.43			
PROX	A38	91.99	95.07	96.95	96.24	100.00	96.23			
PROX	A3	90.33	93.85	96.48	95.43	96.23	100.00			
PROX	A18	90.59	94.02	96.67	95.71	96.55	97.39			
PROX	A52	90.50	94.32	95.81	95.88	96.44	96.84			
PROX	A19	92.27	95.14	96.50	96.84	97.89	96.38			

Figure 3. Example of similarity matrix. The values in column A and row A do not have to be necessarily correlative, as the number indicates the identification value of each site, corresponding to the Sites table.

[2] The conversion and automated taxation of the data were developed by the Salzburg's research team members Slawomir Sobota and Xavier Casassas.

1.1 Visual representation of the data

Whether using a structured database according to Salzburg's methodology (see point 1) or employing simmilarity matrixes (see point 2) multiple functions and tools can be applied in VDM. For instance, different dialectometrical types of visualization are available in numerical and graphical forms (Goebl 2006). The latter, can be viewed as Working, Choropleth, Density, Synopsis, Honey Comb and Beam maps. Dendrograms are also available using Cluster Analyses with diverse algorithms (Single linkage, Complete linkage, Simple average linkage, Average linkage or UPGMA, Centroid, and Ward). Although the method employed depends totally on the researcher's questions and objectives, Complete linkage is considered the best method for geolinguistic analysis and for the interpretation of linguistic trees (Goebl 1993)[3].

Thus, after preparing the data (as a database or simmilarity matrix), the next step is the elaboration of Polygon maps which are based in the geographic references of the enquiry points. Polygon maps can not be made in VDM, since this is a previous task that has to be performed with an adequate Geographic Information System (GIS) software. At Salzburg DM, one of the basic softwares is MapInfo, which is sometimes employed with the software Vertical Mapper[4]. After searching for the exact geographical coordinates of the inquiry points and having processed the geographical projection corresponding to the geographical area of the sampled sites, the map is displayed and polygonized according to the Voronoi/Delaunay method. The coordinates and map data must be then exported to the VDM database where the rest of the information is recorded (see point 1). This exportation process is carried out through the software Meta2Polidef.exe, which has been especially developed using Salzburg DM for this task. With this programme the coordinates exported to the database are linked to every enquiry point.

As a summary, two previous tasks must be performed in order to be able to work with VDM: a) Prepare the information in an appropriate database or elaborate the

[3] However, Ward is also regarded by some DM researchers as one of the most effective mathematical algorithm when dealing with these sort of data, mainly because it minimises squared error (Heeringa and Nerbonne 2001).
[4] See *Desktop Mapping* (2002).

similarity matrixes, and b) Proceed with the cartographic elaboration according to GIS standards.

2. Methodology

Having previously mentioned the methods employed for the proper visualization of the data, we shall now introduce our revision of Viaplana's work.

Unfortunately, original data were not available. However, as similarity matrixes are included in Viaplana's work, we decided to adapt them as it is precisely required for VDM. We chose the first four matrixes of the original work to be reassessed. This included to have a complete squared similarity matrix[5] where both a horizontal row and a vertical column indicated the inquiry points of the map. The matrix had to be saved as a TAB file and then converted into DAT extension with a text file processor (see Figure 4). Once this task was done, the DAT files were opened and saved as a VDM extension.

1	ROWTYPE_		VARNAME_		A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13	A14	A15	A16	A17	A18
2	PROX	A1	100	43	41	55	64	50	56	62	58	48	39	41	45	47	68	62	45	57		
3	PROX	A2	43	100	59	53	46	50	38	40	40	54	51	63	45	51	44	36	47	36		
4	PROX	A3	41	59	100	55	42	50	34	36	36	48	53	55	43	47	40	36	45	35		
5	PROX	A4	55	53	55	100	56	60	48	52	50	52	47	53	47	51	56	52	47	47		
6	PROX	A5	64	46	42	56	100	57	63	63	63	57	42	48	48	54	65	59	42	59		
7	PROX	A6	50	50	50	60	57	100	55	51	55	59	42	54	52	58	53	51	46	57		
8	PROX	A7	56	38	34	48	63	55	100	67	71	55	38	42	44	50	59	67	38	67		
9	PROX	A8	62	40	36	52	63	51	67	100	69	51	38	40	42	46	59	69	38	65		
10	PROX	A9	58	40	36	50	63	55	71	69	100	55	40	42	44	50	61	69	40	68		
11	PROX	A10	48	54	48	52	57	59	55	51	55	100	52	58	56	62	53	51	50	55		
12	PROX	A11	39	51	53	47	42	42	38	38	40	52	100	49	39	43	42	40	39	47		
13	PROX	A12	41	63	55	53	48	54	42	40	42	58	49	100	47	53	42	38	41	40		
14	PROX	A13	45	45	43	47	48	52	44	42	44	56	39	47	100	67	48	44	51	45		
15	PROX	A14	47	51	47	51	54	58	50	46	50	62	43	53	67	100	52	46	55	47		
16	PROX	A15	68	44	40	56	65	53	59	59	61	53	42	42	48	52	100	63	50	58		
17	PROX	A16	62	36	36	52	59	51	67	69	69	51	40	38	44	46	63	100	42	66		
18	PROX	A17	45	47	45	47	42	46	38	38	40	50	39	41	51	55	50	42	100	43		
19	PROX	A18	57	36	35	47	59	57	67	65	68	55	47	40	45	47	58	66	43	100		
20																						

Figure 4. Verbal constituents similarity matrix adapted from Viaplana's work into VDM format. Column A and row A indicate the site location in the map

[5] A distance matrix can also be employed for other purposes. As Goebl (2007) points out 'By a simple transformation (similarity values + distance values = 100) an appropriate distance matrix can be calculated from the similarity matrix'.

Parallel to this, it was necessary to make the Polygon map of the sampled area. Only after all these previous steps were completed, we were able to process a new VDM project and apply all the possible functions and tools of the software into the Northwestern data. During this first stage, we performed a Cluster analysis, obtaining the corresponding dendrograms from three different algorithms: Complete linkage, UPGMA, and Ward. Our results were compared with those obtained in Viaplana's work, where he employed UPGMA clustering method.

3. Results

Our results are completely similar to Viaplana's work when we apply the same algorithm, say, UPGMA, and all the sites in the study are distributed in the same way. This shows that, at least using this clustering method, both analyses are in agreement and the matrixes adapted and processed here with VDM have the same branch ordering. However, Complete linkage and Ward present a slightly different structure of the dialect branches (see Table 1 and the dendrogram demonstration in Figures 5-8).

	Viaplana´s study	*Our study*		
	UPGMA	UPGMA	COMP. LINK.	WARD
Verbs	18	18	8	4
Pronouns	18	18	16	18
Demonstratives & Possesives	18	18	17	18
Phonology	18	18	5	8

Table 1. Nº coincidences in sites distribution using three clustering analysis

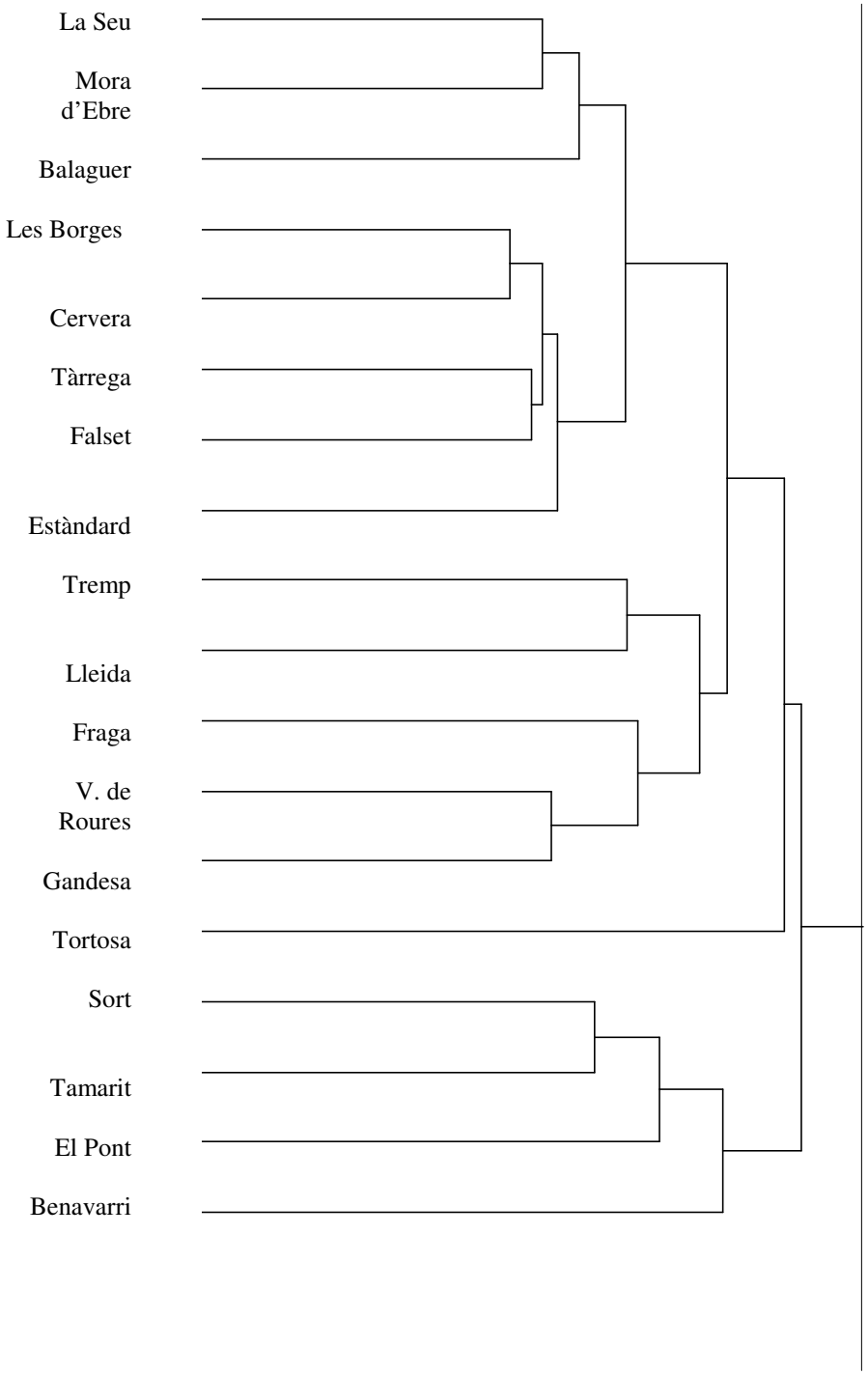

Figure 5. Dendrogram for Verbal constituents using UPGMA method in Viaplana's work

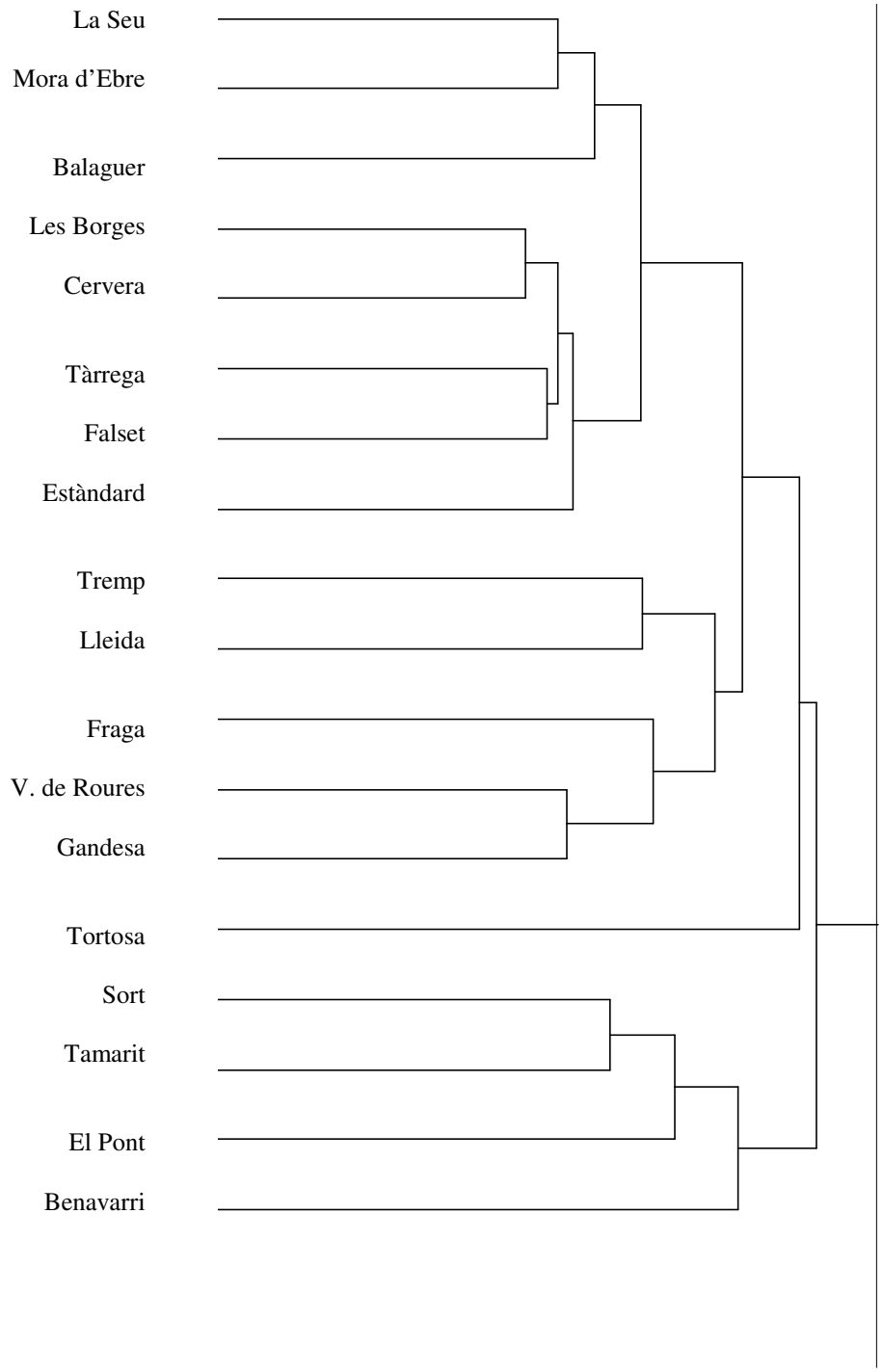

Figure 6. Dendrogram for Verbal constituents using UPGMA method in our study

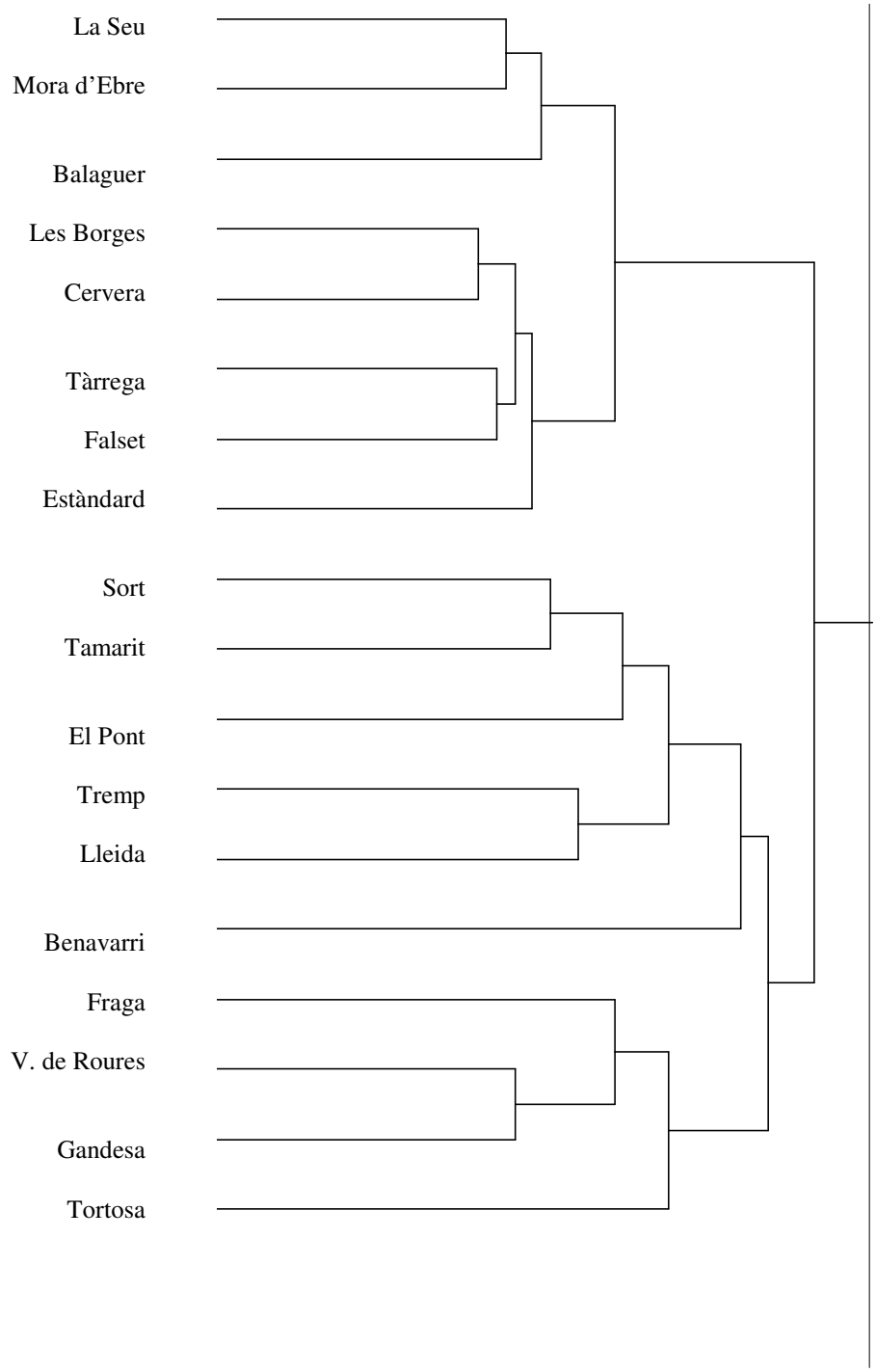

Figure 7. Dendrogram for Verbal constituents using Complete Linkage method in our study

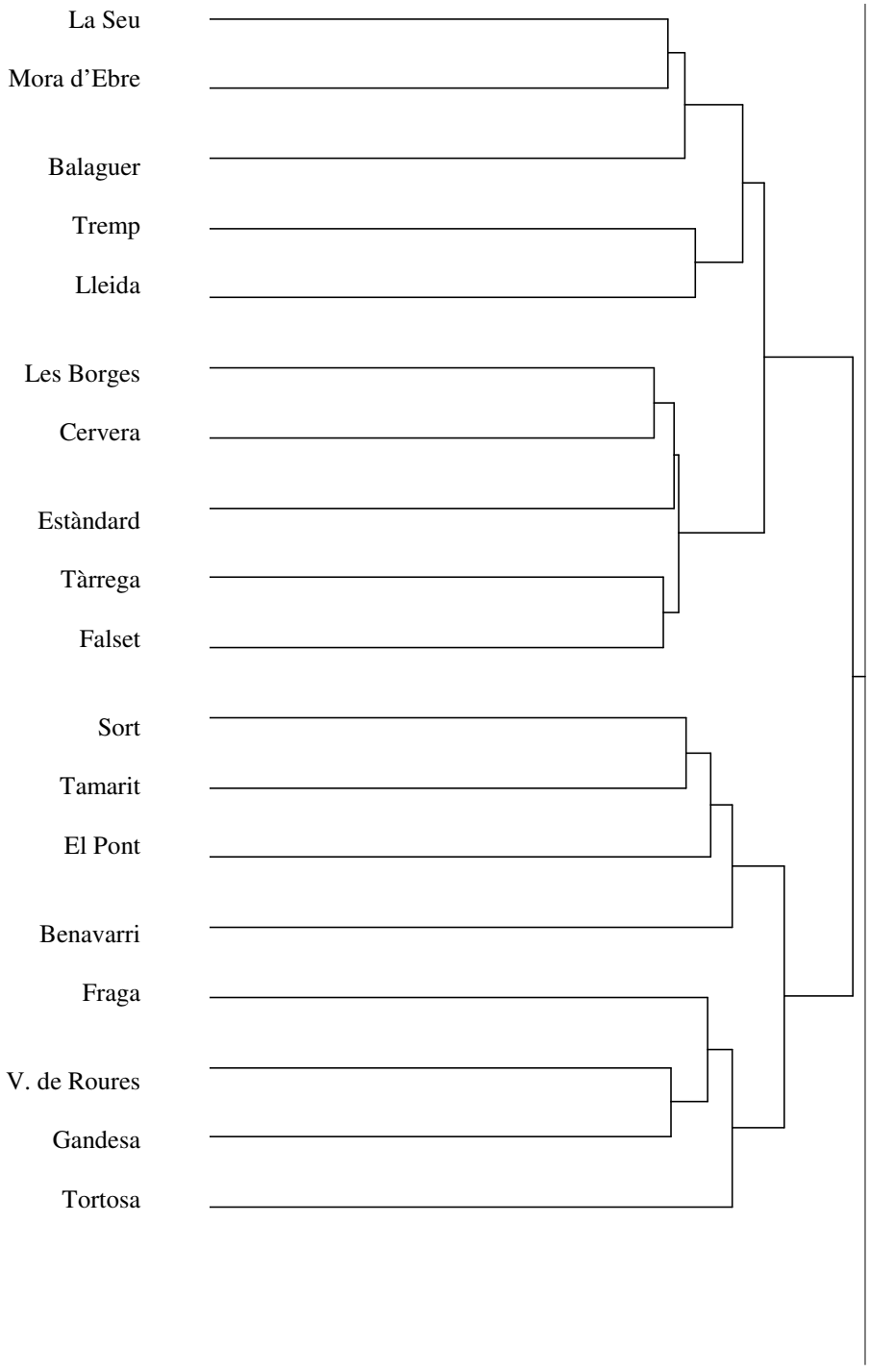

Figure 8. Dendrogram for Verbal constituents using Ward's method in our study

4. Conclusions

We introduce in this paper some basic guidelines to elaborate a VDM project and intend to be in this way a contribution for researchers who might want to employ VDM and learn about the first stages of preparation that have not been mentioned in previous articles regarding S-DM methodology.

As an example of such a project, we present here the results obtained after having applied the tool VDM into data published by Viaplana et al. (1999) on the linguistic distance of the Northwestern Catalan varieties. Viaplana's work represents a great development in the application of dialectometric methodology into the Catalan language and – even though the study is limited to the Northwestern Catalan linguistic domain – his analyses contribute with fresh and new perspectives into the studies of dialect variation in this language.

Thus, we present here our results on clustering analyses obtained from Viaplana's data using the Salzburg's methodology and the software VDM. Our dendrograms show complete similarity to those in the original work when using the same cluster analysis, UPGMA. Differences can be noticed when processing the data with Complete linkage or Ward's method. These new results, which are based in the distribution distinctions found in the trees, might be of great interest to complement Viaplana's study and take a further step in Catalan dialectology investigations. However, future research is necessary in order to explain these differences and account for new approaches that might improve our knowledge in this subject.

5. References

ALCOVER, Antoni & Francesc de Borja MOLL (1929-1932) *La flexió verbal en els dialectes catalans,* Barcelona: Anuari de l'Oficina Romànica de Llengua i Literatura [Published in separate periods during 1929-1932: Vol. 2 (1929) [73] 1- [184] 112, Vol. 3 (1930) [73] 1- [168] 96, Vol. 4 (1931) [9] 1- [104] 96, Vol. 5 (1932) [9] 2 - [72] 64]].

GOEBL, Hans (1993a) "Dialectometry. A Short Overview of the Principles and Practice of Quantitative Classification of Linguistic Atlas Data", in Köhler, R. and Rieger (eds.), *Contribution to Quantitative Linguistics,* Dordrecht/Boston/London: Kluwer, 277-315.

GOEBL, Hans (2006) "Recent Advances in Salzburg Dialectometry", *Literary and Linguistic Computing*, Oxford University Press, Vol. 21, No. 4, 411-435 .

GOEBL, Hans (2007) "A bunch of dialectometric flowers: a brief introduction to dialectometry", in Klein, H., Markus, M. and Schendl, H. (eds.), *Tracing English through time. Explorations in language variation*, vol. 95, Wien: Braumüller, 133-171.

HEERINGA, Wilbert & John NERBONNE (2001) "Dialect Areas and Dialect Continua", in David Sankoff, William Labov and Anthony Kroch (eds.), *Language Variation and Change,* New York: Cambridge University Press, vol. 13, 375-400.

OLBRICH, Gerold; Michael QUICK; Jürgen SCHWEIKART (2002) *Desktop Mapping,* Berlin: Springer.

PEREA, Maria Pilar (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 edition.

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