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

Marcela J. Rivadeneira\textsuperscript{1} and Xavier Casassas\textsuperscript{2}

\textsuperscript{1} Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra
marcela.rivadeneira@upf.edu

\textsuperscript{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 \textit{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 \textit{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 its manyfold instruments of analysis developed on the job (ranging from the similarity maps to correlative DM)'.

VDM employs Salzburg’s DM methodology 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 structuring 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.

---

\(^1\) The first version of VDM was developed by Edgar Haimerl, Hans Goebl’s senior research assistant.
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.

Every invited researcher may use the working methodology system that s/he considers the most appropriate. 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 dialectometrical can use different similarity measures, s/he has to select the appropriate index, keeping in mind her/his theoretical hypotheses on dialectal similarity' (Goebel 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.

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

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 similarity 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 (Goebel 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 (Goebel 1993)³.

Thus, after preparing the data (as a database or similarity 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 Mapper⁴. 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.

<table>
<thead>
<tr>
<th>SIMILARITY (TAB)</th>
<th>VDM (DAT)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>A2</td>
</tr>
<tr>
<td>0.75</td>
<td>0.85</td>
</tr>
</tbody>
</table>

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’.

©Universitat de Barcelona
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).

<table>
<thead>
<tr>
<th></th>
<th>Viaplana’s study</th>
<th>Our study</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>UPGMA</td>
<td>UPGMA</td>
</tr>
<tr>
<td>Verbs</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Pronouns</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Demonstratives &amp; Possesives</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Phonology</td>
<td>18</td>
<td>18</td>
</tr>
</tbody>
</table>

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


