Computer modeling of innovations relative to Latin in today's Romance dialects

Philippe Boula de Mareüil[1], Marc Evrard[1], Alexandre François[2], Antonio Romano[3]

[1] Université Paris-Saclay, CNRS, LISN
[2] Lattice, CNRS, ENS-PSL, Université Sorbonne Nouvelle
[3] Università degli studi di Torino, LFSAG

Europe is characterized by great dialectal diversity, which still needs to be documented. For more than a century, Aesop's fable "The North Wind and the Sun" has been used by the International Phonetic Association to illustrate many languages and dialects spoken in the world. On this basis, a speaking atlas of the regional languages of France was designed, before being extended to other European countries (Boula de Mareüil *et al*. 2018, 2021). The linguistic atlas, available at https://atlas.limsi.fr, allows visitors to hear and read this one-minute story in hundreds of versions, in minority languages or dialects. Most speakers of the atlas, recorded in the field, also translated a list of a hundred words (in particular referring to fauna and flora) into their varieties. Based on these digital data, and following the principles of dialectometry (Nerbonne *et al*. 2007; Patriarca *et al*. 2020), the comparative method and especially historical glottometry (François 2014; Kalyan & François 2018), we propose computational tools to address the following questions. Is there more variation between northern and southern Romance dialects, or between the west and east of the domain (Ibero-Romance dialects on the one hand, Gallo- and Italo-Romance dialects on the other)? How can we quantify it? To what extent do the groupings depend on the levels considered (phonetic, morphosyntactic, or lexical)? This study relies on a sub-corpus illustrating several dozen Romance dialects from France, Belgium, Switzerland, Italy, Spain and Portugal, for which 148 innovations relative to Latin have been encoded as a matrix of 1s and 0s. We encoded:

- phonetic innovations, e.g., regular sound change E > [wa];
- morphological innovations, e.g., merger of Latin imperfects in -ABA- and -EBA-;
- syntactic innovations, e.g., non-null subject or narration in the present perfect;
- lexical innovations, e.g., substitution of CUM 'with' with APUD HOQUE > *avec*.

A range of classification techniques was applied to visualize the emerging clusters (in the form of trees, projections into a two-dimensional plane, etc.) and draw the main isoglosses. The results of the different methods of analysis and calculation will be confronted in order to propose a synthesis, making it possible to reassess the location of the main dividing lines between dialect groups. These results shed new light on dialectology, contributing to model the dynamics of territorial expansion since the breakup of Latin.

Several attribute selection algorithms have been used, among which decision trees provide a readily readable representation. Agglomerative hierarchical clustering (also known as Ward's method) provides other representations, in the form of dendrograms. The Python libraries Scikit-learn and Plotly were used. The former offers various attribute selection and classification algorithms, while the latter enables the results to be visualized in the form of variable-size points and choropleths (i.e., colored surfaces, which may correspond to our dialectal areas). Attribute selection is all the more important as some machine learning algorithms do not allow the number of features to be greater than the number of observation vectors (here, survey points). As some selection algorithms, such as Random Forests, are not deterministic, they were run 100 times, and a majority vote was applied to determine the best

features. From the entire set of features, the survey points considered can be ranked according to the number of innovations they show, and this number can be associated with a more or less dark color on a base map in GeoJSON format and then vectorized.

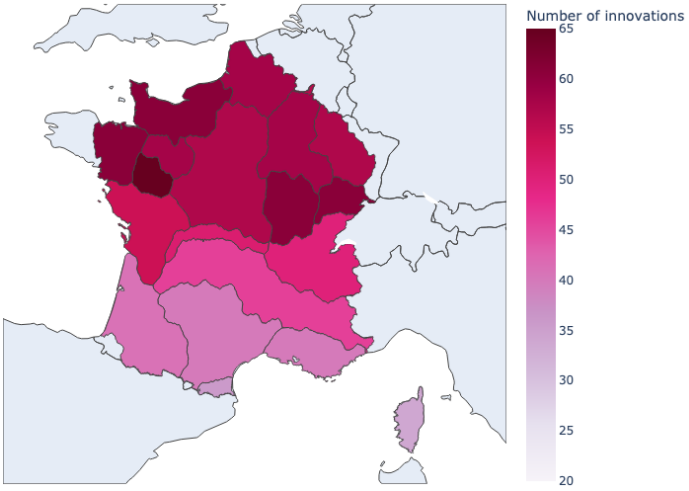

Fig. 1: Linguistic map of France featuring the Romance areas that are the most innovative (in dark red) vs. the least innovative (in light pink).

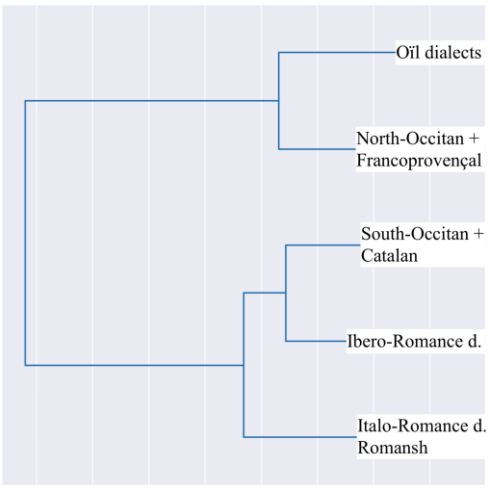

Fig. 2: Simplified dendrogram resulting from a cluster analysis with 5 classes.

The first results show that *Oïl* (northern Gallo-Romance) France is the most innovative (Fig. 1). This is the case, in particular, of the Angevin dialect, in the northwest of the domain, while in France, Corsica is the region with the fewest innovations compared to Latin. At the scale of Europe, Friulian and Romansh (Rhaeto-Romance goup) are the most conservative, according to our measures. Among the most discriminant features, we find the palatalization of Latin CA, which characterizes the majority of Northern Gallo-Romance dialects. The results of the hierarchical clustering, using all features or only the best ones to guarantee the robustness and parsimony of the approach, provide heuristic answers to the questions raised above. Preliminary results (Fig. 2) corroborate a North/South division, the divide passing through the middle of the Occitan (southern Gallo-Romance) area, with *Oïl* and intermediate dialects clustering together in the North, the rest in the South. In the second branch of the dendrogram, the main division is between South-Occitan, Catalan and Ibero-Romance dialects on the one hand, Rhaeto- and Italo-Romance dialects on the other.

References

Boula de Mareüil, P., Rilliard, A., Vernier, F. (2018), A Speaking Atlas of the Regional Languages of France, *11th International Conference on Language Resources and Evaluation*, Miyazaki, 4133–8.

Boula de Mareüil, P., É. Bilinski, V. De Iacovo, R. Romano, F. Vernier (2021), For a mapping of the languages/dialects of Italy and regional varieties of Italian. In A. Thibault, M. Avanzi, N. Lo Vecchio, A. Millour (Eds.), *New Ways of Analyzing Dialectal Variation*. Strasbourg: Éditions de linguistique et de philologie, 267–288.

François, A. (2014), Trees, Waves and Linkages: Models of language diversification. In C. Bowern & B. Evans (eds.), *The Routledge Handbook of Historical Linguistics*, London: Routledge, 161–189.

Kalyan, S. & A. François (2018), Freeing the Comparative Method from the tree model: A framework for Historical Glottometry. *Senri Ethnological Studies*, 98: 59–89.

Nerbonne, J., P. Kleiweg, W. Heeringa, F. Manni (2007), Projecting dialect differences to geography: bootstrap clustering vs. noisy clustering. In C. Preisach, L. Schmidt-Thieme, H. Burkhardt, R.Decker (Eds.), *Data analysis, machine learning, and applicationsy*, Berlin: Springer, 647–654.

Patriarca, M., Heinsalu, E., Léonard, J.-L. (2020), *Language in Space and Time*, Cambridge: CUP.