A measure for heterogeneity in spatial language variation

An entropy-like measurement method for spatial distribution of language items

Schönberg Andreas[1] & Alfred Lameli [2]

[1]PhD student at the department Deutscher Sprachatlas, Philipps University of Marburg, Germany
[2]Professor at and Director of the department Deutscher Sprachatlas, Philipps University of Marburg, Germany

Abstract

Dialectometric studies usually ask about the internally consistent groups of dialects within a language area (see Goebl 1984). However, when dealing with larger sets of geographically specified language data, the problem arises of identifying those regions that are particularly prone to variation or particularly sensitive to language change. The question then is not so much about stability in an area (typically indicated by the definition of clusters), but about instability. More recent dialectometric studies have introduced a number of solutions to this problem, for example, based on resampling techniques (see, e.g., Wieling & Nerbonne 2015). In our project, we follow an approach based on the concept of entropy (e.g., Prokić & Nerbonne 2008) that, in contrast to other studies (Prokić et al. 2009), is not applied to strings of tokens, but geographic distributions.

Our study deals with data from a historical language survey of German dialects at 2500 sites in the regions of Baden (Germany) and Elsass (France). These data are interesting from the perspective that they contain information on different age groups and thus enable analyses on language change (= apparent time; in contrast, analyses in real time become possible by comparison with both more traditional and more recent surveys in the same region).

In order to identify areas which are more sensitive to language change than others we use an entropy-like measure for the identification of heterogeneity/uniformity in spatial language distributions. More concrete, we use a nearest neighbor approach resulting, first, for every linguistic variable of our corpus (e.g., morphemes, lexemes) in a normalized global index with higher values indicating a more homogeneous spatial distribution and lesser values indicating a variative state. We use this global measure for the automated detection of linguistic items with higher/lesser language variation.

Furthermore, a transformation into a local measure of spatial variation makes it possible, second, to automatically identify individual regions with particularly high language variation (typically the transition zones between areas of linguistic variants). This is used, for example, to predict language change or to test the correlation of spatial variation that occurs for different linguistic phenomena. Applying this measure to a collection of multiple linguistic phenomena leads to a new perspective on the structuring of linguistic space highlighting not so much the clusters of linguistic similarity, but the zones of particular linguistic dynamics. The paper will introduce the measure and discuss some examples.

References

Goebl, Hans (1984): Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF. Tübingen: Niemeyer.

Prokić, Jelena, John Nerbonne, Vladimir Zhobov, Petya Osenova, Kiril Simov, Thomas Zastrow & Erhard Hinrichs (2009): The computational analysis of Bulgarian dialect pronunciation. In: Serdica. Journal of Computing(3), 269-298.

Prokić, Jelena and John Nerbonne (2008): Recognizing Groups among Dialects. In: International Journal of Humanities and Arts Computing(2), 153-172.

Wieling, Martijn & John Nerbonne (2015): Advances in Dialectometry In: Annual Review in Linguistics 1(1), 243-264.