Discovering Dialect Regions in Syntactic Dialect Data

Erik Tjong Kim Sang, Meertens Institute, The Netherlands, erik.tjong.kim.sang@meertens.knaw.nl

Arguments for separating one language or dialect from another are often based on historical and political motives. We would like to know if similar linguistic regions can be identified from linguistic data only. For this purpose, we analyze linguistic data from different locations, divide the locations over different regions based on their linguistic similarity and compare the results with traditional dialect boundaries.

Data

We use linguistic data from the Syntactic Atlas for Dutch Dialects (SAND) [2]. These data are based on interviews with people from 267 locations in The Netherlands and Flanders, the Dutch-speaking part of Belgium. The interviews have been transcribed and checked for the presence of predefined syntactic variables. We use the subset of 220 syntactic variables which are available from the digital version of the atlas: DynaSAND [1].

Unlike previous studies that used the SAND for modeling [4, 5], we do not use the evidence of absence assumption to infer additional negative data values but rather keep the unknown values as such in the data set. We also do not split multi-valued variables into multiple binary valued ones because that increases the weights of these values in the model. Instead, we keep multi-valued variable values in the data.

Methods

We used k-means clustering [3] for dividing the 267 locations in different regions. We define the distance between two basic values as 0 when they were equal and 1 when they are different. The distance between sets of values is 0 when the sets share a common value and 1 otherwise. The distance between an unknown value and any other value (including unknown) is defined as 0.5. We define the distance between two locations as the sum of the distances between their syntactic feature values. The two most distant locations were used as initial centers of two abstract regions. The k-means algorithm computed the members of the regions, computed new centers for the regions and repeated itself. After two runs the algorithm stabilized with center locations Erica and Bevere. The resulting geographical map can be found in Figure 1 (left).

Discussion

The clustering algorithm roughly divided the Dutch-speaking community along the national borders: The Netherlands vs Flanders, with the exception of 13 Flemish cities from the province Limburg which were sensibly put in the same region as their dialect neighbor, the Dutch province of Limburg. This result is impressive because no geometric data was used for the classification. As a comparison, a two-dimensional summary of the algorithm's concept of syntactic space, can be found in the right part of Figure 1.

A reasonable notion of linguistic region must present in the data to enable the clustering algorithm to produce the map in the left part of Figure 1. This suggests interesting topics

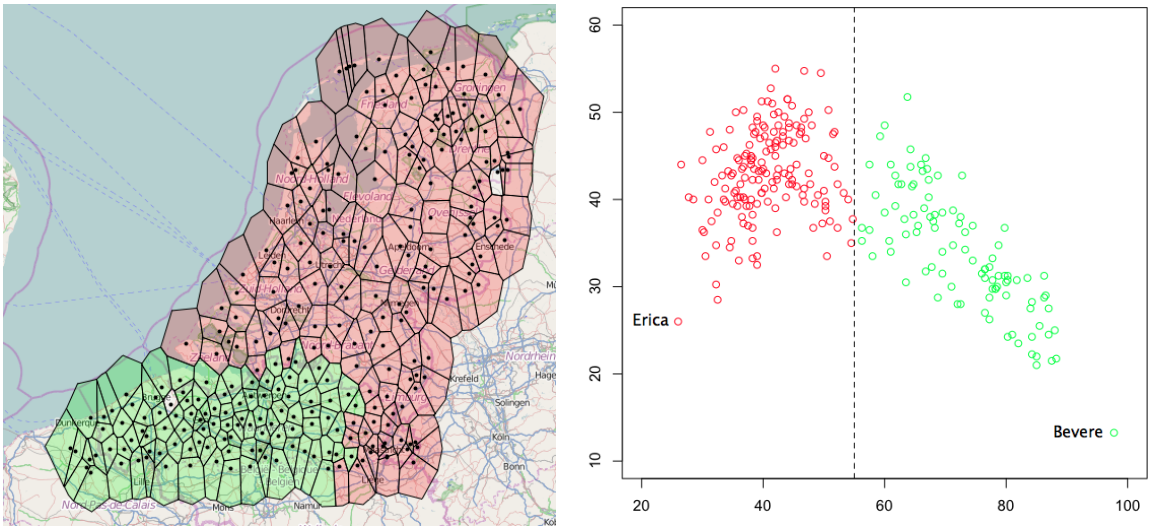

Figure 1: **Left**: A geographic map with the two main Dutch-speaking regions identified by the k-means algorithm from the syntactic SAND data. The two blank locations are the linguistic centers of the two regions: Bevere for Flanders (green) and Erica for The Netherlands (red). **Right**: An abstract map based on the linguistic distances between the locations and the two centers. Each circle represents a location. The Dutch locations (to the left of the central vertical line) are more tightly clustered than the Flemish ones, which indicates a higher syntactic variance of the Flemish dialects.

for future study. Can we derive reasonable smaller dialect regions with this approach? Is the algorithm able to detect smooth transitions between dialect regions, so-called transition zones? Will the method be able to cope with additional input data from a different domain, like morphological data, perhaps even from a different location set?

References

[1] Sjef Barbiers. Dynamische Syntactische Atlas van de Nederlandse Dialecten (DynaSAND), 2006. http://www.meertens.knaw.nl/sand/ Accessed 26 February 2015.

[2] Sjef Barbiers, Johan van de Auwera, Hans Bennis, Eefje Boef, Gunther Vogelaer, and Margreet van der Ham. *Syntactic Atlas of the Dutch Dialects*. Amsterdam University Press, 2008.

[3] Christopher D. Manning and Hinrich Schutze. *Foundations Statistical Natural Language Processing*. MIT Press, 1999.

[4] Marco René Spruit. *Quantitative perspectives on syntactic variation in Dutch dialects*. LOT, Utrecht, The Netherlands, 2008.

[5] Jeroen van Craenenbroeck. The signal and noise in Dutch verb clusters – A quantitative search for parameters, 2014. Manuscript, http://jeroenvancraenenbroeck.net/s/paper_signal_noise.pdf Version 18 December 2014, Retrieved 26 February 2014.