Dealing with ‘non-geographic’ lexical heterogeneity when deriving dialect maps from lexical data

Dirk Speelman & Dirk Geeraerts (University of Leuven)

In dialectometry several approaches exist to automatically generate dialect maps on the basis of calculated dissimilarities between locations (using survey data). Examples for Dutch are Heeringa (2004), who calculates dialect maps on the basis of aggregated phonetic dissimilarities, and Spruit (2008), who calculates dialect maps on the basis of aggregated syntactic dissimilarities. In both examples the automatically generated maps bear a strong resemblance to reference dialect maps compiled by expert dialectologists. However, obtaining similarly successful results on the basis of lexical information has proved to be far more difficult, in spite of the existence of a long research tradition and the availability of authoritative, well studied lexical dissimilarity measures (most notably the gewichteter Identitätswert, Goebel 1984).

One reason for this relative lack of success may be the failure to acknowledge and identify specific ‘non-geographic’ sources of variability in the lexical data. Recently Speelman & Geeraerts (2007) have demonstrated, on the basis of Dutch dialect data, that naming heterogeneity can be modeled successfully in a multiple regression analysis with heterogeneity as the response variable and a series of features of the named concept as the predictor variables (the predictor variables quantified concept salience, concept vagueness and negative affect towards the concept). In this paper we claim and demonstrate that the results from Speelman & Geeraerts (2007) can be put to use in the generation of dialect maps.

We introduce two criteria to assess and compare the quality of different dialect maps and we demonstrate that by enriching the calculation of lexical dissimilarities with a weighing schema which ‘teases apart the aforementioned concept related variability from purely regional lexical variation’ we can improve the quality of the dialect maps according to both criteria. Also, according to both criteria, the resulting dissimilarity measure outperforms the ‘state of the art’ gewichteter Identitätswert (Goebel 1984).

References