Corpus-based dialectometry: Why and how

BENEDIKT SZMRECSANYI
KU Leuven

CHRISTOPH WOLK
University of Giessen

Classical dialectometry draws on linguistic atlas material as its primary data source. By contrast, the bulk of the variationist socio(linguistic) literature adopts corpus-linguistic methodologies. These rely empirically on collections, so-called ”corpora”, of naturalistic, machine-readable texts (for example, interviews) to make claims about linguistic phenomena and/or linguistic variation. Thus unlike other methodologies in linguistics - e.g. those that rely on experimental data, or on elicited linguistic knowledge, or on intuitions (either the linguist's own or somebody else's) - corpus analysis is the methodological outgrowth of the usage-based turn in linguistics. Against this backdrop, our point of departure is that the ability to analyze naturalistic corpus data is central to the maturation of the dialectometry endeavor. The challenge is that corpus-derived dialectological datasets are noisier and dirtier (i.e. less balanced) than atlas- and survey–derived dialectometric datasets. With this in mind, we sketch two ways of conducting corpus-based dialectometry: a top-down approach and a bottom-up approach. The top-down approach first defines a feature catalogue, then establishes frequencies of, or probabilities associated with, these features, and subsequently calculates a joint measure of pairwise linguistic distance between the dialects considered. In bottom-up corpus-based dialectometry, features are not defined a priori, but are allowed to emerge in a data-driven fashion via the identification and subsequent aggregation of significant n-gram patterns. The case studies that we use to illustrate these approaches summarize work by Szmrecsanyi (2013) and Wolk (2014) based on data from the Freiburg Corpus of English Dialects (see www.helsinki.fi/varieng/CoRD/corpora/FRED/).