The future of dialect data - Representation and analysis

JELENA PROKIC
MICHAEL CYSOUW
JOHANN-MATTIS LIST
Philipps-Universität Marburg

Ever since the appearance of the first linguistic atlases in the XIX century (Wenker, 1877; Gilerion, 1880), it was evident that linguistics atlases are more than a sheer collection of the maps that document certain linguistic phenomena. Even in days of relatively modest cartographic technology, researchers have been using various solutions to represent multiple linguistic phenomena simultaneously on a single map and to connect linguistic and different kinds of extra-linguistic information in their atlases (Lameli, 2010). Yet, the developments of information technologies in the past few decades have made it much easier to manipulate larger amounts of data, statistically analyze it, connect it to non-linguistic phenomena, and map it in a single- or multilayer fashion.

In this talk we present new ideas to enhance the representation and analysis of dialect data, and illustrate our ideas with help of a new digital preparation of dialect data that was collected in the 1970s as part of the *Phonetischer Atlas von Deutschland* (PAD) project at the *Forschungszentrum Deutscher Sprachatlas* in Marburg.

Based on the PAD dataset we will illustrate how new tools for quantitative language comparison, which are available as part of a free Python library (LingPy, List & Moran 2013), can be used as a basis for quantitative *data analysis*. We use these tools to automatically segment phonetic transcriptions, and to produce multiple alignments of all cognate words in our data. Although common in evolutionary biology, multiple alignments are still in their infancy in linguistics. In our talk, we will try to illustrate their incredible value for comparative enterprises in dialectology. Once data is available in multiple alignment format, both macroscopic analyses of whole data sets (”aggregate analyses”) as well as microscopic analyses of individual word positions and their phonetic environments can be carried out in a straightforward, fast, and easy way. Furthermore, multiple alignments provide numerous possibilities to produce both qualitative and quantitative geographic maps of feature distributions.

The data is made available in a *linguistic datapackage format*, a format specifically designed for easy data exchange. A *datapackage* is a simple zipped directory in which human-readable text files contain the linguistic data, the relevant geographic data required for dynamic map production and additional metadata. We will argue that this format is much easier to handle than traditionally used XML-files or database-dumps.