Metadata for Analyzing Geolinguistic Variation

Sheila Embleton
York Universtiy
Dorin Uritescu
York Universtiy
Eric S. Wheeler
York Universtiy

Traditionally, dialects are defined by the geographical distribution of certain linguistic variables. With the advent of digital data sets, and modern online dialect atlases, it becomes possible to select a wide range of such linguistic variables, based on the results of *any* search one can make over the relevant data set. The result is a plethora of overlapping dialect regions, with boundaries that can become obscured by the multitude of possibilities.

However, a more nuanced approach, in which one uses variables that are of a common type (such as phonetic, morphological or syntactic) has provided clearer pictures of the dialect areas – even if the pictures vary from one type of data to another.

We have explored this approach, using our digitalized data sets and our online dialect atlas software. To enable the approach even further, we have extended our data sets and software to include metadata about the data from the field. Linguistic features (such as gender on nouns, or tense on verbs) and contextual features (such as the age of the respondent) are coded in the data set with "metatags". Searches over the data sets can then be restricted according to the metatags. Analysis (such as multidimensional scaling or MDS) can be readily directed to different levels of strictly defined linguistic domains.

As a result, we are able to weigh the role of different sets of linguistic features in the dialect fragmentation of the dialect area. We illustrate this aspect by applying MDS to distinct nominal and verbal systems, such as masculine and feminine noun plurals, periphrastic verbal constructions and auxiliaries. Metadata proves to be a valuable step forward in understanding geolinguistic variation.