Isoglosses and distributions of features –
Analyses of the *Dialectological Atlas of the Russian Language*

The Dialectological Atlas of the Russian Language (abbreviated DARJa based on its Russian title) represents more than 4 decades of data collection, from 1938 onwards, and was published in Moscow during 1986-2005. It contains 313 maps, each corresponding to a linguistic feature, and covers 4196 locations. In 2015-16, researchers at Kazan Federal University extracted linguistic features and their values directly from the physical maps and created Excel files giving the values of features across locations (Isaev et al. 2016). We have processed these materials further, georeferencing the map of locations covered and manually extracting the latitude and longitude location of every location.

Thus, the DARJa data are now amenable to systematic, quantitative analyses. For instance, it is possible to define dialect areas in more principled ways than was hitherto possible. For instance, Zaxarova and Orlova (2004: 166) present a map of 28 dialect zones. A rather similar map can be generated from the DARJa data by computing Hamming distances based on the features present in different locations, classifying the locations in a UPGMA tree and cutting this tree into k = 28 clusters. This approach is principled and also versatile, since k can be any number.

The focus in this talk is on two issues of dialectological method, namely how to draw isoglosses computationally and how to measure the similarity of two distributions of feature values. After having binarized all features by taking each feature value as present/ absent we extract isoglosses, as follows. First, we fit a thin plate spline (Franke 1982) to each binarized feature, and produce a spatial interpolation on the region in question (following Wieling et al. 2011 and Guzmán Naranjo and Becker 2021). An example of this can be seen in Figure 1. This figure shows the spatial distribution of so-called *Akan'e* (weakening of unstressed *o*). If the probability of the interpolation is rounded to 0 and 1 we are left with clearly delimited regions that allow for extracting the isoglosses by applying an edge detection procedure to the map. Finding features which have a similar spatial distribution is also performed using the interpolated values. Here we use correlation distances between values across locations. Using these methods for drawing isoglosses and comparing distributions we go on to analyze the interplay between the many phonological, morphological, syntactical, and lexical features in DARJa.

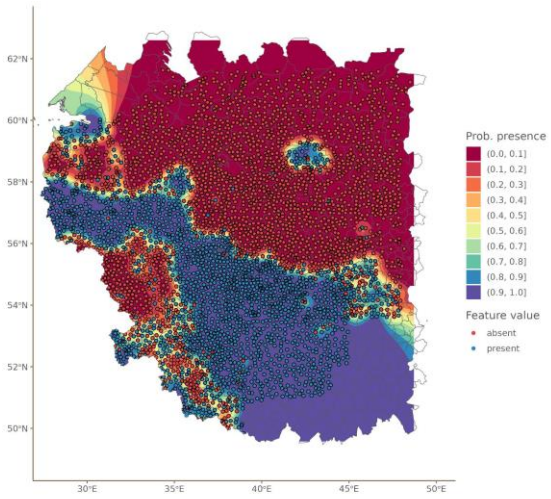

Figure 1: Map of interpolated probabilities for *Akan'e* (weakening of unstressed *o*). The color scheme shows a probability P of presence = 1 as blue vs. P = 0 as red, with the intermediate color range representing interpolated values.

References

Franke, Richard. 1982. Smooth interpolation of scattered data by local thin plate splines. *Computers & Mathematics with Applications* 8.4, 273–281.

Guzmán Naranjo, Matías & Laura Becker. 2021. Statistical bias control in typology. *Linguistic Typology* 26, 605–670. https://doi.org/10.1515/lingty-2021-0002

Isaev, Igor', Valerii Solov'ev, Farid Salimov, Aleksandr Piljugin & Venera Bairaševa. 2016. Creating a database of Russian dialects and prospects for dialectometric studies. *Herald of the Russian Academy of Sciences* 86(6): 448-453.

Wieling, Martijn, John Nerbonne & R. Harald Baayen. 2011. Quantitative social dialectology: Explaining linguistic variation geographically and socially. *PloS ONE* 6, e23613.

Zaxarova, Kapitolina F. & Varvara G. Orlova. 2004. *Dialektnoe členenie russkogo jazyka*. Moskva: URSS.