# Materials in the DARJa repository

## 1. Data and data preparation

The R file darja_data3.RData, which contains an object called d, has information on geographical IDs, geographical coordinates, feature type, feature, and feature values. A corresponding tab-delimited text file with the same rows is offered solely for those who do not wish to engage with R; it's called darja_data3.txt.

Files used in the preparation of darja_data3.RData are: DARJa_data_preparation2.R, data-small-final2.csv, Cintaksis.xlsx, Fonetika.xlsx, Leksika.xlsx, Morfologiya.xlsx. The last four were prepared by a team at Kazan Federal University and downloaded by us on 2022-03-30 (they are still available at https://kpfu.ru/atributivnaya-baza-dannyh-russkih-govorov-269324.html as of 2025-07-21). The Kazan files have IDs for locations, coordinates in an unidentified system, and features and their values. Using the script darja_data3.RData we extracted this information and combined it with information on geographical coordinates gathered by ourselves through georeferencing the generic DARJa map—information contained in data-small-final2.csv. None of the files DARJa_data_preparation2.R, data-small-final2.csv, Cintaksis.xlsx, Fonetika.xlsx, Leksika.xlsx, Morfologiya.xlsx are used in subsequent analyses, but they are supplied in order to document the nature of the original data and details of the data extraction.