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, and is given as a zip file.

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. Note that locations with the IDs 628 and 2329 do not have proper numbers on the DARJa maps, and their map numbers are given as 0 or -2 in the Kazan files, even if the files do have features and feature values associated with them in those files. They are considered spurious and were deleted from our data, which then has 4193 rather than 4195 locations.

2. Computing linguistic distances

The R file linguistic_distance_matrix4.RData, which contains an R object called m, contains linguistic distances among the 4193 locations. Row and column names correspond to the IDs, so they are sequences from 1 to 4195 with 628 and 2329 missing. The distances were computed using linguistic_distances_darja2.R. That script also contains a toy example for demonstration of how the calculation works. Because of constraints on the size of files at GitHub, access to the linguistic distance matrix is provided as a link in the file distance_matrices_links.txt.

3. Computing geographical distances

The R file geographical_distance_matrix3.RData, which contains an R object called mgeo, contains geographical distances among the 4193 locations. Row and column names correspond to the IDs, so they are sequences from 1 to 4195 with 628 and 2329 missing. The distances were computed using geographical_distance_matrix.R. Because of constraints on the size of files at GitHub, access is provided as a link in the file distance_matrices_links.txt.

4. Clusterization and cluster validation

We produced ultrametric trees based on the distance matrix, using all methods available in the hclust() function of base R, except 'single', 'WPGMC', and 'UPGMC', which put nearly all locations in one big cluster even when k = 5, for instance. The method used was: 'ward.D', 'ward.D2, 'complete', 'average' (same as UPGMA), and 'mcquitty' (same as WPGMA). The trees are in the files WardD.RData, WardD2.RData, complete.RData, UPGMA.RData, WPGMA.RData, and the objects contained in the files have the same names as the files minus the suffix (so: WardD, WardD2, etc.). Clusterization was carried out by clusterization_darja.R.

Experiments with cluster validation were directed at finding some quantitative criteria for deciding between the 5 methods and some optimal number of clusters. All experiments are carried out using optimal_number_clusters4.R. The first section of the script implements using the Adjusted Rand index (ARI) to quantify the fit between the distribution of feature values over locations (dialects) and a given clusterization. Output of this is in optimal_number_clusters2.txt. The second section does silhouette scores on the clusterizations. The third section does stability scores, output of which is in stability_scores.txt. A final section contains some qualitative notes made during inspection of clusters in their geographical setting using the dmap() script (for which see below).

5. Rivers

Using rivers.R, the files rivers_m.txt and rivers_l.txt were produced. The script and the text files, as zipped files, are provided. The files list all pairs of locations that are less than 300 km with their linguistic and geographical distances, indicating the number of rivers or lakes that separate them. The river data come from https://www.naturalearthdata.com/downloads/ via the interface of the rnaturalearth R package. The file name elements rivers_m and rivers_l refer respectively to the medium and large scale data of naturalearth. 'Medium' means 1:50 m(illion) (medium resolution, relatively few rivers) and 'large' 1:10million (higher resolution, relatively many rivers). The data in rivers_m.txt and rivers_l.txt are read and analyzed using analyze_river_data.R.

6. Biomes

The script biomes.R reads biome data from Olson & Dinerstein (1998) posted at https://www.worldwildlife.org/publications/terrestrial-ecoregions-of-the-world#:~:text=There%20are%20867%20terrestrial%20ecoregions,conserve%20biodiversity%20around%20the%20world. The files are in https://files.worldwildlife.org/wwfcmsprod/files/Publication/file/6kcchn7e3u_official_teow.zip. Unzipped, this will yield a folder, which should be renamed to official_teow and be placed as a subfolder of the folder where biomes.R is found. The script biomes.R identifies the biome associated with each location in DARJa and creates a data frame with IDs, coordinates, and a number corresponding to the biome of each location. This is output to the file biomes_locs.txt. The numbers are associated with descriptions in the little file called key_biomes98.txt. Subsequently, biomes.R creates a tab-delimited file called biomes3.txt with a row per pair of locations giving the location IDs,

the geographical and linguistic distances, and a 0 for ‘same’ and 1 for ‘different’ biome. This large file is shared via a link in the biome_data_link.txt; its contents is further analyzed using the script analyze_biome_data.R. This script investigates the effect of biomes on linguistic distance when controlling for geographical distance. Results are in biome_results.txt and the biome_difference.jpeg plot.