Beyond example extraction: Quantitative analysis of the JANES corpus

Maja Miličevič*

* Faculty of Philology, University of Belgrade, Serbia

Short description

The goal of the workshop is to provide an introduction to quantitative analysis of corpus data using the R environment. The rationale is that (1) quantitative analysis is needed to properly describe corpus data, and in particular to generalise from one language sample to other similar samples and language in general; (2) R is one of the most powerful tools for quantitative analysis out there, and is freely available. The workshop will be divided in three sessions, dedicated in turn to basic considerations of corpus data and R, sample description and statistical inference. All sessions will use (meta)data from JANES.

Prerequisites: Experience in work with corpora will be assumed. No previous knowledge of statistics or R is required; an introductory handout will be provided about a week before the workshop to help participants brush up some basic math concepts and form expectations about R.

Session 1: "Obtaining data from corpora: How and why?"

- introduction to quantitative corpus studies
- formulating linguistic hypotheses testable on corpus data
- the R environment: installing R, setting working directory, installing packages
- importing data into R: defining and coding variables, file formats

Session 2: "Describing and visualising corpus data"

- descriptive statistics: counts, frequency distributions; mean, median; standard deviation, interquartile range
- graphs: scatter plots, line charts, bar charts, histograms, box plots

Session 3: "Generalising from corpus data"

- basics of statistical hypothesis testing: intro to probability; null hypothesis, significance levels and their meaning; parametric vs. non-parametric statistics
- some specific tests: chi-square, correlation, (intro to) regression