Count your frequencies wisely! An introduction to concepts and methods in quantitative (corpus) linguistics

Maja Miličević (University of Belgrade) and Adriano Ferraresi (University of Bologna)

28 October 2016, LabTerm (Via G. della Torre 1, Forlì)

Description

The goal of this one-day workshop is to provide an introduction to quantitative analyses in linguistics using the R environment, with a special focus on the analysis of corpus-derived data. The rationale is that a firm grasp of quantitative methods is needed to properly describe corpus data, as well as to generalise from a single language sample to other similar samples, and to language in general. The R software is chosen as it is one of the most powerful tools for quantitative analysis, and is also freely available. The workshop's three main sessions will be dedicated to: 1) fundamental concepts in the design of corpus-based studies and the basics of R; 2) description of samples, and 3) statistical inference. A practice session will follow. Though the main focus will be on corpus data (obtained from monolingual comparable, parallel and/or intermodal corpora), other data types commonly used in linguistic studies will also be touched upon (e.g. questionnaire and informant data).

Prerequisites

Experience in working with corpora will be assumed. No previous knowledge of statistics or R is required. An introductory handout will be provided to help participants brush up some basic math concepts and form expectations about R.

Schedule

9.00-10.30 Session 1: **"From corpora to R: obtaining and organising data"**

- Introduction to quantitative corpus studies
- Formulating linguistic hypotheses testable on corpus data
- The R environment: installing R, setting working directories, installing packages
- Importing data into R: defining and coding variables, file formats

10.30-11.00 *Break*

11.00-12.30 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, mosaic plots

12.30-14.00 *Lunch break*

14.00-15.30 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
- Specific tests: chi-square, correlation, (intro to) regression

15.30-16.30 Session 4: **"Practice"**

- Hands-on analyses, questions, discussion…