

Multivariate exploratory techniques for linguists with R

Once corpus linguists have collected sizeable amounts of observations and described each observation with relevant variables, they look for patterns in the data. When the data set is too large, it becomes impossible to summarize the table with the naked eye, and summary statistics are needed. This is where exploratory data analysis steps in.

I will introduce four multivariate exploratory techniques: correspondence analysis, multiple correspondence analysis, principal component analysis, and exploratory factor analysis. These techniques rely on an attempt to simplify complex multivariate datasets to facilitate interpretation.

Each method will be illustrated with case-studies that are related to the themes of the TaLC2020 conference. Participants are welcome to bring their own datasets!

Participants will learn how to:

- collect and/or load corpus data into R;
- detect and avoid the common pitfalls of quantitative analysis;
- run multivariate exploratory statistics;
- interpret and report the results in a research paper.

Although no expert knowledge of R is required, a basic knowledge of computers and office software is strongly recommended as we will be working with text files and spreadsheets.

For any questions/queries relating to this workshop, contact: guillaume.desagulier@univ-paris8.fr.