Quantitative and statistical analyses in corpus linguistics:
a practical introduction

Wed 23 May: 14:00-17:15

Tutors:
Stefan Evert
Stefan Th. Gries
Sebastian Hoffmann

Draft Programme

1) Foundations [14:00 - 15:00]

- the need for statistical analysis in corpus linguistics
- sampling variation and basic hypothesis tests (one-sample case)
- the meaning of p-values and significance levels
- effect size and confidence intervals
- frequency comparison (two-sample case)
- how to choose the size of a random sample

2) Examples with BNCweb [15:00 - 15:30]

- the main focus is how to cast a linguistic research question as a quantiative problem and apply statistical methods
- frequency comparison procedure illustrated with BNCweb data
- per million words or per thousand sentences? - the case of relative clauses with *whom*
- extrapolating from manually checked / annotated random samples: sampling strategy and sample size

coffee break [15:30 - 16:00]

3) Visualization and simple stats with the computer [16:00 - 16:45]
(see http://www.linguistics.ucsb.edu/faculty/stgries/teaching/icame2007/index.html)

- loading statistical tables in R
- life without the random sample assumption (where your data points are relative frequencies for each document rather than whole-subcorpus frequencies)
- visualisation of data for correlations and means
- parametric and non-parameteric correlation coefficients
- non-parametric tests for means (U-test and Kruska-Wallis ANOVA)

4) Surgery [16:45 - 17:15]