Lecture 4. Lexico-grammar: From simple counts to complex models

Aim: The lecture is based on Brezina (2018), Chapter 4. It focuses on the statistical analysis of lexico-grammatical features in language such as articles, passive constructions or modal expressions. The chapter shows how lexico-grammatical variation can be summarised using cross-tabulation and what statistical measures can be computed based on cross-tabulation summary tables. These measures range from simple percentages to the chi-squared test and logistic regression.

- Key terms: outcome variable | predictor | lexico-grammatical frame | cross-tabulation | mosaic plot | chi-squared test | log likelihood test | probability ratio | logistic regression | parsimonious model | odds ratio

Time:

1-hour lecture.

2-hour computer lab session with exercises and Lancaster Stats Tools online (optional).

1-hour individual study (readings).

Statistical tools: Crosstab, Categories comparison and Logistic regression.

Practical exercises: Chapter 4 Exercises and answers.

Data: Chapter 4 data.

Readings: Brezina 2018 Chapter 4 + optionally selected Advanced readings recommended in the book.

Outline:

1. Analyzing a lexico-grammatical feature
2. Crosstabulation, percentages and chi-squared test
3. Logistic regression

Main points – 'Things to remember':

- When analysing lexico-grammatical variation we need to pay attention to individual linguistic contexts and define a lexico-grammatical frame.
- Cross-tabulation can be used for a simple analysis of categorical variables. In addition to frequencies, crosstab tables can also include percentages based on row totals (most useful for investigation of lexico-grammar), column totals and the grand total.
- The data in cross-tab tables can be effectively visualized using mosaic plots.
- We can test the statistical significance of the relationship between variables in a two-way crosstab table (i.e. a table with one linguistic and one explanatory variable) using the chi-squared test. The effect sizes reported are Cramer's V (overall effect) and probability or odds ratios (individual effects).
- Logistic regression is a sophisticated multivariable method for analysing the effect of different predictors (both categorical and scale) on a categorical (typically binary) outcome variable.
- In logistic regression, we look at both the general performance of a model as well as at individual coefficients showing the effect of the predictor variables on the outcome of interest.