Lecture 5. Register variation: correlation, clusters and factors

Aim: The lecture is based on Brezina (2018), Chapter 5. It discusses a group of methods that can be used for the simultaneous analysis of a large number of linguistic variables that characterise different texts and registers. First, we look at the relationship between two linguistic variables by means of correlation. Both Pearson's and the non-parametric Spearman's correlations are explained. Next, we explore the classification of words, texts, registers etc. using the technique of hierarchical agglomerative clustering. Finally, the lecture deals with Multi-dimensional analysis (MD), a methodology which uses factor analysis to extract patterns across multiple variables.

- Key terms: correlation, Pearson's correlation, Spearman's correlation, cluster analysis, dendrogram, multidimensional analysis, factor, scree plot

Time:

1-hour lecture.

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

1-hour individual study (readings).

Statistical tools: Correlation calculator, Clusters and MD analysis.

Practical exercises: Chapter 5 Exercises and answers.

Data: Chapter 5 data.

Readings: Brezina 2018 Chapter 5 + optionally selected Advanced readings recommended in the book.

Outline:

1. Relationships between variables: correlation
2. Classification: Hierarchical agglomerative cluster analysis
3. Multidimensional analysis (MD)
4. Things to remember

Main points – 'Things to remember':

- Correlations are used for the investigation of the relationship between two variables at a time.
- Pearson's correlation is suitable for scale variables, while Spearman's correlation assumes ordinal variables (ranks). Spearman's correlation can also be used with scale variables if the means as the measures of central tendency do not represent the data well (extremely skewed distributions).
- Hierarchical agglomerative cluster analysis is used for classification of words, texts, registers etc. The result of this analysis is a tree plot (dendrogram).
- The most complex type of analysis out of the three discussed in this chapter is multidimensional analysis (MD). MD analyses a large number of linguistics variables and reduces them to a small number of factors which are interpreted as dimensions of variation. Along these dimensions, different registers can be placed.