Lecture 1: Introduction: Statistics meets corpus linguistics

Aim: The lecture is based on Brezina (2018), Chapter 1. It introduces basic principles of statistical thinking that are necessary for informed application of statistical procedures to corpus data. It explains the role of statistics in scientific research in general and corpus linguistics in particular. Topics such as the creation of corpora, types of research design, basic statistical terminology, as well as data exploration and visualization will be discussed.

▸ Key terms: statistics sample population descriptive statistics inferential statistics mean sampling frame bias research design visualization

Time:

1-hour lecture.

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

1-hour individual study (readings).

Statistical tools: Stats calculator, Randomizer, Graph tool.

Practical exercises: Chapter 1 Exercises and answers.

Data: Chapter 1 data.

Readings: Brezina 2018 Chapter 1 + optionally selected Advanced readings recommended in the book.

Outline:

1. What is statistics? Science, corpus linguistics and statistics
2. Basic statistical terminology
3. Building of corpora and research design
4. Exploring data and data visualisation
5. Things to remember

Main points ‘Things to remember’:

- Corpus linguistics is a scientific method.
- Successful application of statistical techniques in corpus linguistics depends on the use of a well-constructed unbiased corpus.
- Statistics uses mathematical expressions to help us make sense of quantitative data.
- Effective visualization summarizes patterns in data without hiding important features.
- Although most visible, p-values form only a (small) part of statistics.
- ‘Statistical significance’, ‘practical importance’ and ‘linguistic meaningfulness’ are three separate dimensions which shouldn’t be confused.