Keyword and collocation statistics: Under the hood of CQPweb

Stefan Evert
Friedrich-Alexander-Universität
Erlangen-Nürnberg, Germany
stefan.evert@fau.de

Andrew Hardie
Lancaster University, UK
a.hardie@lancaster.ac.uk

1 Overview

Many corpus analysis programs perform certain statistical calculations “behind the scenes”. The most common such calculations are for keywords and collocations, both of which are based on the calculation of an effect size or a statistical significance test statistic.

The aim of this workshop is to give participants a behind-the-scenes understanding of how these statistics work – taking as our example the CQPweb system’s approach to the calculation of keyword and collocation measures. (These are almost entirely equivalent to how the same measures are treated in BNCweb – see Hoffmann et al. 2008, chapter 6 for details.)

2 Outline

No knowledge of statistics will be assumed. Instead, we will work from the ground up, introducing the three key notions:

- contingency tables, and how they are constructed for keyword analysis and collocation analysis
- how significance test statistics are calculated from such tables
- how effect size measures summarise a quite different quality of the contingency table.

The specific statistics we’ll look at are the following:

- Significance statistics:
    - Chi-squared
    - Log-likelihood
    - Fisher exact test
- Effect size measures:
    - Mutual information
    - Log Ratio (filtered and unfiltered)

Participants will learn to work through these analyses by hand before seeing how they are implemented under the hood in the CQPweb software, of which the workshop leaders are the two main developers. We’ll also introduce the other measures available within CQPweb for collocation specifically, which have often been argued to represent compromises between significance measures and effect size measures – these include the Dice coefficient, the Z-score, and MI3 (“mutual information cubed”).

At the end of the workshop, participants will have a firmer and deeper understanding of the procedures upon which not only CQPweb but many other corpus tools rely.

3 Reference

Hoffmann, S., Evert, S., Smith, N., Lee, D., and Berglund Prytz, Y. (2008) *Corpus Linguistics with BNCweb: a Practical Guide*. Peter Lang.