Statistical Analysis of Corpus Data with R
— Exercise Sheet for Unit #2 —

In this exercise, your task is to extract frequency information from the British National Corpus using the *BNCweb* interface (Hoffmann *et al.* 2008) and to perform statistical frequency comparisons for these data in R.

Participants of a SIGIL course will receive a password for the BNCweb server at

http://www.cogsci.uni-osnabrueck.de/~CL/resources/services.html

Other people can use the public BNCweb demo server at

http://bncweb.lancs.ac.uk/

after applying for a free account here:

http://corpora.lancs.ac.uk/BNCweb/home.html#access

1. Log into one of the BNCweb servers and familiarise yourself with the Web interface and its Simple Query Syntax (CEQL). Learn how to search for word forms, lemmata and phrases, as well as for lexico-grammatical patterns (optional).

2. Pick a word, phrase or grammatical pattern of interest, and calculate its distribution across text types (or other metadata categories). You will have to enter the resulting counts manually into R, unfortunately.

3. Perform frequency comparisons (wrt. words as unit of measurement) for various pairs of categories. Which differences are significant? Do you think that their effect size makes them linguistically relevant?

4. If you perform pairwise frequency comparisons for all text types, you will have to carry out 28 hypothesis tests in total. What could be a fundamental problem of such an approach (apart from being extremely tedious)?

5. The R functions `fisher.test()` and `chisq.test()` can also be applied to a $2 \times n$ contingency table in order to compare all n categories at once. Construct such a table from your data using `cbind()`, `rbind()` or `matrix()`. Is there a significant difference between your categories? What exactly is the null hypothesis of this test?

6. The phrase `{click/V} on` (CEQL query) is significantly more frequent in "other published material" than any other text type. Can you think of a possible explanation for this observation? You might want to take a closer look at the dispersion count (number of different texts) and some corpus examples.

References

Hoffmann, Sebastian; Evert, Stefan; Smith, Nicholas; Lee, David; Berglund Prytz, Ylva (2008). *Corpus Linguistics with BNCweb – a Practical Guide*, volume 6 of *English Corpus Linguistics*. Peter Lang, Frankfurt am Main.