*Marco Baroni & Stefan Evert* http://SIGIL.R-Forge.R-Project.org/

# 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 be given access to a BNCweb server at

https://corpora.linguistik.uni-erlangen.de/bncweb/

Other people can use a public demo server at http://bncweb.lancs.ac.uk/ after applying for a free account (http://bncweb.lancs.ac.uk/bncwebSignup).

1. Log into the BNCweb server 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 can either note down the resulting counts and enter them manually into R, or copy and paste the distribution table displayed by BNCweb into a text editor or spreadsheet software.[1]

3. The BNCweb distribution table includes total word counts for each category. Are word tokens a sensible unit of measurement? If not, use a second query to obtain suitable by-category totals and combine them with the frequency counts from above.[2]

4. Perform frequency comparison tests for various pairs of categories. Which differences are significant? Do you think that their effect size makes them linguistically relevant?

5. 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)?

6. 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, e.g. using `rbind()` to combine two row vectors. Is there a significant difference between your categories? What exactly is the null hypothesis of this test?

7. 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.

[1] Users of Microsoft Excel should make sure to paste the table as plain text rather than HTML.

[2] You may not be able to carry out this step because of data size limitations for your BNCweb account.