Statistical Analysis of Corpus Data with R
— Exercise Sheet for Unit #4 —

In the first part of this exercise, you will practise collocational analysis as explained in the lecture slides by applying the procedure to a new data set based on *surface cooccurrence*. The second part focuses on the application of association measures to *keyword extraction*, searching for words that are particularly characteristic of spoken or written English. The two data sets used for this exercise are included in the `SIGIL` package.

- If you haven't done so already in Unit 1, download and install the `SIGIL` package from CRAN (see installation notes in handout from Unit 1).
- We will use the `BNCInChargeOf` and `BNCcomparison` data sets included in this package. After loading the package with `library(SIGIL)`, familiarise yourself with the data sets by reading the respective help pages (with `?BNCInChargeOf` and `?BNCcomparison`).
- The `BNCInChargeOf` data set contains surface collocates of the phrase *in charge of*, extracted from the British National Corpus. Re-read the description of contingency tables for surface cooccurrences in the lecture slides, then calculate the contingency table of observed frequencies from the provided frequency information (`f.in`, `N.in`, `f.out`, `N.out`). Use `transform()` to add the new variables `O11`, `O12`, `O21` and `O22` to the data set.
- Compute the expected frequencies, row/column marginals, sample size, and association scores for a selection of measures, following the instructions in the lecture slides. Rank the data set according to each association measure. Which measure gives the intuitively most plausible ranking?
- Association measures can also be used to identify characteristic *keywords*, which are much more frequent in spoken than in written English, or vice versa. The data set `BNCcomparison` lists the frequencies of a selection of English words in the written and spoken part of the British National Corpus.
- Construct appropriate contingency tables for the frequency comparison setting, as explained in Unit 2. First, determine the written and spoken sample sizes by summing over all rows of the data set. Then calculate the observed frequencies `O11`, `O12`, `O21` and `O22` for each word (= row), and add them to the data set.
- Which association measures might be sensible for keyword extraction? Compute the respective association scores using the same procedure as above, and rank the data set by *keyness* for written or spoken English. Do high/low association scores correspond to written or to spoken keyness? Compare the keywords identified by different measures. Do you notice any specific problems of individual measures?