

# Statistical Analysis of Corpus Data with R

## Exercise Sheet #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 part of the `corpora` package available from CRAN.

- If you haven't done so already, download and install the `corpora` package from CRAN. On Windows and Mac OS X, use the package installer included in the R GUI. If you are working with R on the command line, experiment with the function `install.packages()` (start by reading the help page, `?install.packages`).
- We will use the `BNCInChargeOf` and `BNCcomparison` data sets included in the `corpora` package. After loading the package (`library(corpora)`), familiarise yourself with the data sets by reading the respective help pages (`?BNCInChargeOf` and `?BNCcomparison`).
- The `BNCInChargeOf` data set contains positional collocates for the phrase *in charge of*, extracted from the British National Corpus. You can load this data set into your R session with the command `data(BNCInChargeOf)`. 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 the lecture on *Hypothesis Testing for Corpus Frequency Data*. 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?