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

Statistical Analysis of Corpus Data with R
— Exercise Sheet for Unit #1 —

In this first exercise, you will familiarise yourself with basic R operations, reading the R documentation, and working with data frames, which are the fundamental statistical data structure used by R. For this purpose, the SIGIL package contains a data frame with complete metadata information for all 4,048 texts of the *British National Corpus*.[1] You should be able to access the BNC metadata under the name BNCmeta after loading the package:

```
> library(SIGIL)
> nrow(BNCmeta)      # check that metadata table is available
[1] 4048
> attach(BNCmeta)    # for convenient access to table columns
```

Here are some things to do with the metadata table. You should try yourself on all these tasks for full credit. Notice that not all the commands are in the course slides: explore the documentation and other R teaching materials.

1. How many rows and columns does the metadata table have? How many meta-information variables are annotated? Read the help page for the data set.
2. How many different genres are there in the BNC? Which genre contains the smallest number of texts? (Hint: use `table()` to obtain frequency counts for the levels of a factor variable.)
3. Save the metadata table to a `.csv` file (recall that CSV stands for *comma-separated values*), which you can load into a spreadsheet application such as Microsoft Excel or OpenOffice.org Calc. Use the spreadsheet to explore the metadata tables.
4. Summarise the distribution of text lengths (measured either in number of words `n_words` or number of sentences `n_s`). The text type *written-to-be-spoken* contains some "outliers", i.e. texts with rather unusual lengths. Use `boxplot()` to identify these outliers. Can you list the titles of the outlier texts?
5. Do text lengths differ between text types? Do they differ between male and female authors? (You can use the "formula" interface to `boxplot()` for this task, but you will have to remove outliers first or rescale the plot.)
6. Later in the course, we might perform machine-learning experiments to distinguish between male and female authors. Produce a subset of the metadata table containing only texts for which author sex is known, omitting the title and irrelevant metadata columns (esp. those which have only a single value in the subset).
7. The number of words and number of sentences as measures of text length should be highly correlated. Illustrate this correlation with a suitable plot. Compute the correlation coefficient between these two variables and its 95% confidence interval. Judging from the plot, do you think there is a simple linear relationship?

[1] See http://www.natcorp.ox.ac.uk/ for more information about the BNC and available metadata.