Pragmatics / text linguistics 2: key words

The examples we will deal with in this topic group will be a little different from the previous ones. It should have become sufficiently clear by now how to generate frequency lists, concordances, and collocate displays so that further examples of such simple applications are not required anymore. It would therefore be pointless to have you write yet another small script to generate a concordance for one or more words where the only difference is that this time you generate them for a pragmatic / text-linguistic purpose. Examples for such studies include Okada's (1999) study of respective and respectively based on data from the BNC, Oh's (2000) corpus-based study of actually and in fact in the Brown corpus and the Switchboard corpus, Aijmer's (1984, 1986) and Gries and David's (to appear) study of the hedges kind of and sort of (based on the BNC), etc.

The pragmatics / text linguistics case studies will therefore be slightly different. Instead of concordancing etc., they are concerned with two approaches to text-linguistic methods that are available in concordancing software and that we would, therefore, like to be able to do with R, too. On the one hand, this would once more show you that you do not need commercial programs, and in fact we will see how R can again excel over such programs; on the other hand, it will broaden your horizon because you will put R to use in ways you probably haven't thought about yet.

One application of frequency lists involves the comparison of frequency lists from two corpora in order to identify the words that are (significantly) more frequent in one corpus than in the other (reference) corpus. This can be used to compare different linguistics varieties or sociolects. For example, Hofland and Johansson (1982) study word frequencies in British and US English, using the Brown corpus and the LOB corpus to identify words that are more typical of one variety of English than the other. They use the chi-square test as well as Yule's difference coefficient. Johansson and Hofland (1989) improve on this work by figuring in word class information. Leech and Fallon (1992) adopt a very similar approach to investigate what words which are significantly more frequent in British or American English reveal about cultural differences. Rayson, Leech, and Hodges (1997) use the conversational part of the BNC and conduct chi-square tests to determine word-frequency differences between female vs. male speakers, as well as between speakers from different age and social groups. Rayson and Garside (2000) use the log-likelihood statistic to compare a target corpus of air traffic control communication against a reference corpus (a part of the spoken data from the BNC). Oakes (2003) does very much the same as the early studies yet again; the only differences are that he uses two other corpora, namely FROWN and FLOB, and he very briefly also tests another method, high ratio pairs. All these approaches can be implemented with R with what you know.

In most of the cases, the corpus to be investigated is a smaller and/or specialized corpus while the reference corpus is a larger and/or more balanced corpus. In our present example, we will compare two texts of nearly identical sizes.
Assignment

Write a script that has the following characteristics and performs the following operations:

(i) The script prompts the user to choose the files `<C:/_qclwr/inputfiles/corp_perl.txt>` (as a specialized corpus) and `<C:/_qclwr/inputfiles/corp_python.txt>` (as a reference corpus), which contain the texts of the Wikipedia articles on the programming languages Perl and Python (as downloaded on April 12, 2006).

(ii) The script loads the input files and splits them up into words at the occurrences of at least one non-word character.

(iii) The script accesses each word type attested in at least one corpus at least once and
− generates a 2×2 table containing the frequencies of that word type in both files compared to the summed frequencies of all other word types in both files;
− computes a chi-square test (with correction) for each word type and sets this chi-square value to a positive value when the word is preferred in the Perl corpus and to a negative value when the word is preferred in the Python corpus;
− saves each word, the data from its 2×2 table, its expected frequencies, and its chi-square value into a data frame called result that is sorted in descending order according to the keyness of the word types for the Perl file.

A more comprehensive version of this script would not just record the word type and its (positive or negative variant of the) chi-square value but also the frequencies $a$, $b$, $c$, and $d$ for future processing. You may wish to explore this possibility, too.

One problem remains with this approach. Many of the 2×2 tables will violate the assumption of the chi-square test mentioned in Section 5.2, namely that 80 percent of all expected frequencies are larger than five, which is why, depending on what your script looks like, you may get warnings. Strictly speaking, it is therefore better to use a different test; two powerful alternatives are the log-likelihood test and the Fisher-Yates exact test. The formula for the log-likelihood test statistic $G^2$ is given in (1), where $ln$ refers to the natural logarithm (i.e., the logarithm to the base $e$, which is 2.7182818):

\[
G^2 = -2 \sum_{i=1}^{4} \text{obs. frequency} \cdot \ln \left( \frac{\text{obs. frequency}}{\exp(\text{frequency})} \right)
\]

(1)

Maybe you should try to revise the script accordingly, again setting $G^2$ values to positive values when the word is preferred in the Perl corpus and to negative values when the word is preferred in the Python corpus. If you decide to do so, include the frequencies $a$, $b$, $c$, and $d$ in your output. In R, you get the natural logarithm of a number $n$ by typing `log(n)` or, more explicitly, `log(n, base=exp(1))`. Note, however, that `log(0)` gives `-Inf`, which if multiplied by 0 yields `NaN` (meaning "Not a Number"). Thus, if you observe observed frequencies of 0, set the corresponding summand(s) to 0 manually.

A final possibility for exploration is to have the script output or plot the coefficient of difference used by Hofland and Johansson (1982) and Leech and Fallon (1992). It is defined as in (2) and if you do it cleverly, you can do this with about two lines of code …

\[
difference \text{ coefficient} = \frac{\text{Freq word } w \text{ in corpus}_1 - \text{Freq word } w \text{ in corpus}_2}{\text{Freq word } w \text{ in corpus}_1 + \text{Freq word } w \text{ in corpus}_2}
\]

(2)
When you are done, load the files \texttt{<C:/qclwr/scripts/prag-text\_2\_assignment\_chi.r>} and \texttt{<C:/qclwr/scripts/prag-text\_2\_assignment\_ll.r>}, \texttt{<C:/qclwr/outputfiles/prag-text\_2\_result\_chi.txt>} and \texttt{<C:/qclwr/outputfiles/prag-text\_2\_result\_ll2.txt>}, as well as \texttt{<C:\qclwr\outputfiles\prag-text\_2\_assignment-plot.png>}, and compare your solution with them.

For further study/exploration: the studies mentioned before the assignment and Manning and Schütze (2000:175f.) for relative frequency ratios, a similar descriptive measure which does not involve significance testing.