Chapter 10 – Corpus Linguistics in Cyberspace

10.1 In Table 10.1 in the book, frequency figures for web searches for different from, different than and different to at different points in time are reported.

(i) Go to Google, click on “Advanced search” and type in different from in the search box for “exact phrase”. Make the search and note the frequency figure given.

(ii) Repeat for different than and different to.

(iii) Compare your figures with those in Table 10.1. Do the figures indicate that the web has continued to grow? Do the phrases seem to increase at the same rate?

10.2 Some words in English can either be spelt -ise or -ize as in specialise/specialize. Some people consider -ize to be American and -ise to be British, while in fact the -ize variant is now the first option in most British dictionaries as well. Many British publishers and writers, however, stick to the -ise spellings. Investigate the distribution of -ise/-ise spellings on the web.

(i) Go to Google, click on “Advanced search” and type in realise normalise globalise in the search box for “any of these words”. Under “Domain”, type “uk”. Make the search and note down the frequency.

(ii) Repeat (i) for the following domains: edu, au, nz, za, hk, in and ph.

(iii) Repeat (i) and (ii), but with the search words spelt realize normalize globalize.

(iv) Fill in the figures in a table like Table 10.3 in the book. Can you see any patterns in the distribution of these spellings?

10.3 For some investigations, a linguistic interface to the web can be advantageous. In this task we will use WebCorp Live to study the use of the words catastrophe and disaster in British broadsheets (quality newspapers).
(i) Go to http://www.webcorp.org.uk/. Click on “Advanced Search Options”. Type in the word catastrophe in the search box. Set “Case options” on “Case insensitive”. Under “Output Format”, choose “HTML Tables (KWIC)”. Scroll down to “Site domain/Country” and click on “UK Broadsheet Newspapers”. Then click on “Submit” at the bottom of the page and wait for the results. WebCorp is a bit slower than the other interfaces you have used, so have patience!

(ii) When you get the results, scroll down to the bottom of the page. There you will find a dialogue box with “Advanced Options”. Go to “Sort Concordances Alphabetically”. To the left, scroll down and choose “left side”. For position, choose “1”. This will sort the concordance lines alphabetically on the word immediately left of the keyword. Click on “Submit”.

(iii) Go through the sorted results and note down the most frequent adjacent collocates of catastrophe.

(iv) In the “Advanced Options” dialogue box, go to “Collocation” and click on “External Collocates” and “Exclude Stopwords”. This will calculate the most frequent collocates of catastrophe, but exclude the most common function words like a, the etc. Then click on “Submit” and wait for the program to work.

(v) A collocates table will be shown above the “Advanced Options” dialogue box. Here you can see the most frequent collocates, and also in what position they occur. Copy the table and paste it into a word processing document.

(vi) Repeat (i)–(v) for the word disaster.

(vii) Compare the results for catastrophe and disaster. Are the two words synonyms? Are they used about the same kind of problems?

10.4 Study some other phenomenon by means of WebCorp Live. Try changing some of the settings. Note that you can get more tokens by changing the number of pages from 100 up to a maximum of 500.