Chapter 2 – Counting, Calculating and Annotating

2.1 In Table 2.3 in the book, we can see which fifteen noun lemmas are most frequent in the BNC. In this exercise we are going to investigate whether the same noun lemmas occur at the top of the list in COCA.

(i) Go to http://www.americancorpus.org/. NB: during the session you may be asked to register as a user. Do this – it is free of charge. Choose the same password as you chose for BYU-BNC.

(ii) Find out the frequencies in COCA for the top fifteen noun lemmas in Table 2.3 in the book. You will have to search for one word at a time, using the following method.

(iii) In the search panel to the left, click “List” and type in the following search string: [time].[nn*]. The square brackets around the search word will retrieve all forms of the lemma (in this case time and times). The second part of the string, [nn*], ensures that only nouns will be retrieved. Note that there are no spaces between the brackets and the period in the middle. Since in this task we are only interested in the relative rank of the nouns, there is no need to calculate tokens per million words: absolute frequencies will do.

(iv) Go through the procedure in (iii) for each of the fifteen top words in Table 2.3 and note down the frequencies. Then order the nouns according to frequency and compare your list based on COCA with the list based on the BNC. Are there large or small differences? What conclusions can you draw from the results?

2.2 Individual words or lemmas are usually not evenly distributed in a corpus. For instance, Table 2.10 in the book shows for a set of adjectives that their frequency per million words varies considerably from one subcorpus to another. Make a similar investigation of a set of verbs, viz. argue, cook, dispute, speak and talk.

(i) Log in to COCA.

(ii) In the search panel, click “Chart” and type in the following search string: [argue].[v*]. Remember that the square brackets around the search word retrieves all forms of the lemma, in this case argue, argued and arguing, while [.v*] makes sure that only verbs are retrieved. Then click on “Search”.

(iii) You will now get the results in a bar chart. Create a table like Table 2.10 in the book in your word processing program or on paper and write in the results for argue in the columns.
for Spoken, Fiction, Magazines, Newspapers and Academic. For this task, disregard the figures given for different time periods to the right in the bar chart.

(iv) Repeat with the other four verbs.

Were the words distributed as you had expected? Were there any surprises?

2.3 Table 2.11 in the book shows that in British English the simplex preposition out is more frequent in spoken language, while the complex preposition out of is more frequent in written language. Investigate whether the same holds for American English in COCA, by checking two common phrases: out (of) the door and out (of) the window.

(i) Log into COCA. In the search panel, click the “List” button and type in the following search string: out of the window/door. Then under “Section” choose “Spoken”. Then click “Search”. The numbers for out of the window and out of the door will be shown in the right-hand panel. Create a table like Table 2.11 in your word processor or on paper and fill in the total figure for “Spoken” under out of.

(ii) Next, search for the same string in the written subcorpora. In order to search in Fiction, Magazine, Newspaper and Academic at the same time, hold down the shift key as you click on them one at a time. Continue as in (i) and note down the total figure for “Written” under out of in your table.

(iii) Repeat the procedure in (i) and (ii) for out the window/door and fill in the totals in the table.

(iv) Now calculate the row percentages. First add the figures for out and out of in the “Spoken” row. Then divide the figure for out by the total and multiply by 100: that will give you the percentage for out. The percentage for out of will simply be the rest up to 100. Fill in the percentage figures in your table. (For example, if your figure for out had been 60 and for out of 300, you would have made the following calculations: 60 + 300 = 360; 60:360 = 0.16666; 0.16666 \times 100 = 16.66, which means that the percentage of out would have been 16.66 per cent, which should be rounded off to 16.7 per cent or 17 per cent.)

(v) Repeat the calculations for the “Written” row and fill in the percentage figures.

(vi) Now compare your table with Table 2.11 in the book.
If you have done the task correctly, the American data show a different picture from the British. In what way?