Welcome to the companion for *Corpus Linguistics and the Description of English* by Hans Lindquist and Magnus Levin

Here you will find additional material to be used together with the book. Please report any faulty links or other problems to customerservices@eup.ed.ac.uk.

Last updated: November 2017

Contents

1 Links to online corpora and other resources used in the exercises

2 Web-based exercises for each chapter in the book
   - Chapter 1 Corpus linguistics
   - Chapter 2 Counting, calculating and annotating
   - Chapter 3 Looking for lexis
   - Chapter 4 Checking collocations
   - Chapter 5 Finding phrases
   - Chapter 6 Metaphor and metonymy
   - Chapter 7 Grammar
   - Chapter 8 Male and female
   - Chapter 9 Language change
   - Chapter 10 Corpus linguistics in cyberspace

1 Links to online corpora and other resources used in the exercises

A. Corpora available through Brigham Young University
   *(Corpus of Contemporary American English (COCA), Corpus of Historical American English (COHA), British National Corpus (BYU-BNC), Global Web-Based English (GloWbE), News on the Web (NOW), TIME Magazine Corpus)*

   https://corpus.byu.edu
B. Other corpora and resources

AntConc
http://www.laurenceanthony.net/software.html

Michigan Corpus of Academic Spoken English (MICASE)
http://quod.lib.umich.edu/m/micase/

Phrases in English (PIE)
http://phrasesinenglish.org

2 Web-based exercises for each chapter in the book

Chapter 1 Corpus linguistics

1.1 In this exercise you will get acquainted with the British National Corpus (BNC), one of the corpora used throughout the book, and with some of the basic concepts of corpus linguistics. The search interface, BYU-BNC, is identical with the other corpora available through Brigham Young University.

Before you do the web exercise, read about concordances and frequencies in section 1.4 and about the BNC in section 1.6.2.

(i) Go to http://corpus.byu.edu/bnc/. N.B.: during the session you may be asked to register as a user. Do this – it is free of charge. Make a note of your password. Read the site Introduction and then click on ‘five minute tour’ to learn more about the corpus.

(ii) When you are finished with the tour, type in mysterious in the search panel to the left of the screen. Click the ‘Chart’ button above the search word. Click on ‘Section’ (do not tick the box to the left) and mark ‘Ignore’. Now press the ‘See frequency by section’ button. The results per section of the corpus will be displayed in a bar chart on the right-hand part of the screen. The chart shows that mysterious occurs most frequently in Fiction, with 27.2 tokens (instances) per 1 million words, followed by Miscellaneous, Newspaper, Academic and Spoken.

(iii) Now click on the bar for Fiction in the chart (don’t click on the word fiction). The first 100 concordance lines with mysterious will be displayed and you will be able to study the contexts in which the word occurs, e.g. what kind of nouns it modifies.

(iv) Go back to the search panel to the left by clicking in ‘Search’ at the top of the page and mark ‘List’ instead of ‘Chart’. Click on
‘Find matching strings’ again. Now the total number of occurrences is reported in the right-hand panel. If you click on the word *mysterious* you get the first 100 concordance lines regardless of subcorpus.

(v) Look at this list of five adjectives: *big, hot, superfluous, supreme, tragic*. Try to guess which is most frequent, second most frequent etc. in the corpus, and then in which subcorpus they occur most frequently. Then search for each of the words in the manner described above and for each of them note down the total frequency and which subcorpus it occurs most frequently in. Don’t forget to look at the concordance lines to get an indication of how the words are used.

(vi) Try searching for some other words that you are curious about.

**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 https://corpus.byu.edu/coca/. N.B.: 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 and no period after the last bracket. Click on ‘Find matching strings’. 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. Mark down the total for *time* and *times*.

(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, namely argue, cook, dispute, speak and talk.

(i) Log into 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, argues, argued and arguing, while .[v*] makes sure that only verbs are retrieved. Then click on ‘See frequency by section’.
(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 in the bottom part of 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 in contexts like out of/out the window. 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 ‘Sections’ choose ‘Spoken’. Then click ‘Find matching strings’. 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 processing program 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: $60 + 300 = 360; \frac{60}{360} = 0.16666; 0.16666 \times 100 = 16.66\%$, which means that the percentage of *out* would have been 16.66\%, which should be rounded off to 16.7\% or 17\%.)

(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?

Chapter 3  Looking for lexis

3.1  As mentioned in section 3.4, the word *regime* is associated with a certain ‘semantic prosody’ or evaluative colouring, so that it often does not just mean ‘method or system of government’, but rather ‘disagreeable method or system of government’. Take a look at some examples of this word in the Time Corpus.

(i) Go to http://corpus.byu.edu/time/. N.B: 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 and COCA.

(ii) In the search panel, click the ‘Chart’ button and type in the search word *regime*. Then click ‘See frequency by section’.

(iii) The results will be displayed in a bar chart showing the frequency distribution per decade. Click in the bar for the 1920s to see the concordance lines. Study the first twenty lines carefully and decide for each instance of *regime* whether it has negative or positive evaluation or seems to be neutral. If you need to see more context, click on the date to the left of the concordance line. To get back to the concordance from the expanded concordance line, click on ‘Chart’. Write down your results.

(iv) Repeat the procedure for the 2000s.
Did you find any semantic prosody/evaluative colouring in these forty concordance lines? Is there any indication of change from the 1920s to 2000s? (To investigate such a change in meaning properly you would have to study many more examples, but this small sample could give a hint.)

3.2 Section 3.5 describes how words can change in frequency over time, and Figure 3.6 in the book illustrates this with the examples greenhouse effect and global warning in the Time Corpus. Now try similar investigations with some other words in the Corpus of Historical American English.

(i) Go to https://corpus.byu.edu/coha/. In the search panel, click the ‘Chart’ button. Then type in the search words motor car and click on ‘See frequency by section’. The frequency distribution per decade will be given in a bar chart. You can copy and paste the whole chart into a word processing document, or else write down the frequencies per million words. Then repeat the procedure for the following words, one at a time: car, automobile, auto. Report on your findings by making a diagram like Figure 3.6 in the book, or by showing the four bar charts.

(ii) Investigate at least one other set of words with similar meaning from a field that interests you, for instance sports, politics, fashion or arts, or study at least four single words.

Chapter 4 Checking collocations

4.1 When you investigate ‘window collocations’ you study which words occur close to the keyword. Investigate which verbs occur frequently with the word city in COCA.

(i) Log into COCA. In the search panel, click the ‘Collocates’ button and type in the search word city. Under ‘POS LIST’ (short for part-of-speech list), scroll down to ‘verb.ALL’ and choose this. Now _v* will appear in the context field. Change the window size to five words to the left and five words to the right. Then press ‘Find matching strings’. Study the first thirty verbs. What kind of verbs are they?

(ii) Repeat the procedure in (i), but set the ‘SORT/LIMIT’ option to ‘Relevance’. This will give you a list based on mutual information. Compare the verbs you find here with those you found in (i).
(iii) Investigate a few more words from different word classes in the same manner. Try using different size windows and different spans to the left and to the right and see what changes.

4.2 You can investigate adjacent collocations in COCA by using the same method as in 4.1, but narrowing the window size. In this task, investigate what kind of revolutions people speak and write about in COCA.

(i) Log into COCA. In the search panel, click the ‘List’ button, set the ‘SORT’ option to ‘Frequency’ and type in the search word revolution. Under ‘POS LIST’, scroll down to ‘adj.ALL’ and choose this. Now _j* will appear in the context field. Change the window size to one word to the left and zero words to the right to retrieve only the adjectives that occur immediately before revolution. Then press ‘Find matching strings’. Study the first thirty collocates. What kind of adjectives are they? Do they predominantly say where the revolution took place, what was revolutionised or in what manner the revolution was carried out?

(ii) Repeat the procedure in (i), but choose ‘noun.ALL+’ under ‘POS LIST’ to get all nouns that collocate with revolution in the slot immediately to the left. Study the first thirty collocates. What kind of nouns are they? (In fact, you may wonder whether some of these words are nouns at all – some of them, like Republican, can be either nouns or adjectives and may be wrongly tagged as nouns in these examples.)

(iii) Repeat the procedure in (ii), but under ‘SORT’ choose ‘Relevance’. Note that now some collocations with a low total frequency come higher on the list. Compare the two lists. Which do you find more interesting?

Chapter 5 Finding phrases

5.1 In section 5.2, a Google search showed that the idiom storm in a teacup/tempest in a teapot was manipulated in various ways in about every second instance of use. In this task we will investigate another well-known idiom: too many cooks spoil the broth.

(i) Open Google on the internet. In the search box, type in: “too many * spoil the” (don’t forget the quotation marks). Apart from too many cooks spoil the broth, this will retrieve all examples with small modifications like too many chefs spoil the soup and more imaginative
variants like *too many clicks spoil the sale* and *too many superlatives spoil the telling*. Now, under ‘Settings’, click ‘Advanced search’. Here you can limit your search in a number of ways. For this exercise, in the ‘Website or domain’ box, type in ‘uk’ to get examples from British webpages only. Then click on ‘Advanced search’. Among the first thirty examples, how many were manipulated? How were they manipulated?

(ii) Go back to the ‘Advanced search’ page and in the search box ‘without any of these words’ type in: *cook, chef, broth, soup*. Then click ‘Search’ again. This will give you only manipulated examples.

(iii) Try the searches described in (i) and (ii) in some other domain, perhaps your home country if it is not the UK.

(iv) Investigate some other idioms or proverbs of your own choice. You will see that the extent to which they are manipulated varies considerably.

5.2 Many words are frequent because they are part of frequent phrases. Think about the word *ear*. How is it used in English? Is it part of any frequent phrases? To investigate this, use PIE.

(i) Go to http://phrasesinenglish.org. Under ‘Grams’, choose ‘N-Grams Simple’. In the search box, type in *+ + + + ear*. The four + signs make the program retrieve all 5-grams with four words before the keyword. Then change the ‘Minimum frequency’ to 3 and click on ‘Search’.

(ii) Study the list of 5-grams closely. Some may be just chance occurrences, but many will be examples of more or less fixed phrases. Write down the ones you recognise and group those that belong together. If you are uncertain about the meaning and want to see more context, click on the phrase to get the concordance lines (depending on server traffic, this does not always work – if it doesn’t, you can find the same concordance lines via the BYU-BNC interface).

(iii) How is *ear* used in the BNC? In what kind of phrases does it occur?

Chapter 6 Metaphor and metonymy

6.1 Two of the conceptual metaphors mentioned by Lakoff and Johnson (1980) are *LIFE IS A CONTAINER* and *LIFE IS A JOURNEY*. Investigate what ‘life’ is likened to in COCA.
(i) Log into COCA and click on the ‘List’ button. Type in the search string life is a *(note the asterisk) and click on ‘Search’.

(ii) Analyse the first 100 examples given in the right-hand part of the window and try to organise them into semantic groups. How many fit the conceptual metaphors suggested by Lakoff and Johnson? Can you suggest some more conceptual metaphors?

(iii) Now turn to ‘love’. Try searching for love is a *. You will find that many of the words following love is a are adjectives, so you cannot see what metaphor is used. To solve this problem, search for the following string instead: love is a _j*_ _nn*. (You can either type in _j*_ and _nn* or go down to ‘POS LIST’, scroll down and click on the tags you want one at a time.) This will retrieve all examples like love is a winding way and love is a temporary derangement.

(iv) Analyse the love examples in the same way as you did the life ones.

These ways of looking for metaphors are not perfect, since they will not find instances like I want to fill my life with something meaningful or Our love has grown and many others.

6.2 Another way to find metaphors in corpora is to search for words which from experience you know or suspect are used in metaphorical expressions. For instance, see to what extent game of chess is used in metaphors in the BNC.

(i) Log into BYU-BNC. Search for the phrase game of chess. Analyse the examples. How many are (a) references to real chess games, (b) similes and (c) metaphors?

(ii) Repeat the procedure with the words roller-coaster ride. To retrieve all possible spellings, first search for roller*coaster ride, which will retrieve rollercoaster ride and roller-coaster ride, and then search separately for roller coaster ride. Analyse and answer the same questions.

(iii) Think of some other words which are likely to be used in metaphors and metonymy and investigate them in the BNC.

Chapter 7 Grammar

7.1 It is well known that shall is used more in British English than in American English. Compare the frequencies of the two word forms shall and will in COCA and in the BNC.
(i) Log into BYU-BNC. Click on ‘List’. Then click on the greyed ‘POS’, which changes to ‘Insert PoS’ and scroll down to ‘pron. PERS’ and choose that. _pp* should now appear in the search box. Type in the word shall after _pp* and press ‘Find matching strings’. Write down the total figure given at the bottom. Then repeat for the word will and write down the total figure for that.

(ii) Log into COCA (you should be able to stay logged into BYU-BNC in the background). Repeat the settings described in (i). Search for _pp* shall and _pp* will in turn and write down the total figures.

(iii) Now calculate the percentage of shall in the BNC in the following way. Add the figures for shall and will. Then divide the number of shall by the total and multiply by 100. Note down the percentage.

(iv) Repeat the procedure in (iii) for COCA. Compare the two percentage figures. If you have done all the operations correctly, there should be a considerable difference between the two corpora in the proportion of shall.

(v) Now return to COCA. Keep all the settings except ‘List’; change this to ‘Chart’. Search for _pp* shall again. First study the top half of the chart. In which genres is shall predominantly used in American English? Then look at the bottom part of the chart. Is there any tendency to a change in frequency over time?

(vi) Next return to BYU-BNC and repeat the procedure in (v). Can you see any major difference between the corpora in the distribution between the subcorpora?

Chapter 8 Male and female

8.1 Some of the studies referred to in the book focus on the way men and women speak, others on the way men and women are referred to. In the first task we will look at the Michigan Corpus of Academic Spoken English (MICASE) to investigate the way men and women use language.

(i) Go to http://quod.lib.umich.edu/m/micase/.

(ii) Click on ‘Search MICASE’.

(iii) In the search form, click on ‘Female’ under search attributes and type in sorry in the search box. Click on ‘Submit search’.

(iv) Write down the frequency of sorry, which is given at the top of the results page. Note that you immediately get the concordance and can see how the keyword is used.

(v) Repeat the procedure in (iii) and (iv) for the speaker category ‘Male’.
(vi) Repeat the procedures in (iii)–(v) for *pardon* and *excuse me*.

(vii) The statistical overview of the corpus states that the gender proportions in the corpus are 54% female, 46% male. Compare the frequencies of *sorry, pardon* and *excuse me* with those figures. Can you draw any conclusions about male and female language from this? You may want to go back to the concordances to look more closely at how the keywords are used in the discourse situation.

8.2 Think of some other linguistic features that it might be interesting to investigate in MICASE from the point of view of gender. Note that you can specify many more features, such as discourse type and academic role of speaker, but since the corpus is relatively small, if you specify too much the numbers may be quite low.

8.3 Table 8.9 in the book shows the adjectives most frequently used with *man* and *woman* in the BNC. Now make a similar investigation in the BNC and COCA of what items occur in the frame *bis/her + Adj + Noun*.

(i) Log into BYU-BNC. Click on ‘List’. Type *bis* in the search box. Then go down to ‘POS LIST’ and select ‘adj.ALL’ so that _j* appears in the search box. Repeat this action and select ‘noun.ALL’ so that _nn* appears in the search box after _j*. The search string should now be: his _j* _nn*.

(ii) Click on ‘Search’. Among the results, mark the first twenty examples, copy them and paste them into a word processing document.

(iii) Repeat (i) and (ii) but change *his* to *her*.

(iv) Log into COCA and repeat (i)–(iii).

(v) Compare the lists for *his* and *her* in the BNC. What differences can you see? Then compare the lists for *his* and *her* in COCA. Can you draw any conclusions from these results?

8.4 The terminology relating to gender and sexuality has changed over time, because of various social developments. Use COCA to see how some of these terms have changed in frequency over time.

(i) Log into COCA and click on ‘Chart’. Type in the search word *gay* and click on ‘Find matching strings’. Copy the chart into a word processing document, or note down the normalised frequencies for each decade.
(ii) Repeat for the following words: homosexual, lesbian. Note the frequencies and compare with gay.

(iii) Investigate some other individual words or sets of words related to gender or sexuality in a similar manner.

**Chapter 9 Language change**

9.1 We have already seen signs of lexical change in Exercise 8.4, but we now turn to grammar. It has been claimed that the use of the progressive is increasing. Check this in COHA.

(i) Log into COHA and click on ‘Chart’. Go down to ‘POS LIST’, scroll down to ‘verb.ING’ and choose that. _v?g* should now appear in the search box. Press ‘Find matching strings’ and look at the results in the chart. You cannot expect sudden changes in grammar, so the slow but steady rise suggests that progressives have indeed increased in COHA over this period.

(ii) However, we still do not know whether progressives have increased relative to the total number of verbs. To check this, repeat (i), but instead of ‘verb.ING’ choose ‘verb.ALL’. As you will see, the total frequency of verbs per million words has also increased. Compare the rate of change for progressives and all verbs. Do the progressives increase more or less than verbs in general?

9.2 Another area which is believed to undergo change is verb complementation. This exercise concerns the different alternatives found with the verb help.

(i) Log into COHA and click on ‘Chart’. In the search box, type in help to. Then go down to ‘POS LIST’, choose ‘verb.ALL’ and click on ‘Find matching strings’. Copy the chart into a word processing document or note down the changes in frequency.

(ii) Repeat (i), but type in just help and choose ‘verb.INF’ under ‘POS LIST’. Click on ‘Find matching strings’ and save the chart or note the figures.

(iii) Compare and comment on the development of the two constructions, as in help to explain vs. help explain.

(iv) Try to think of other verbs where there is similar variation and investigate whether there is any ongoing change.

9.3 Read about Mair’s (2004) study on start to and start V-ing in section 9.4 in the book.
(i) Make a study of start to and start V-ing in the same way as you studied help to and help + bare infinitive in Exercise 9.2. Compare your findings with Mair’s.

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, and type in different from in the search box. Make the search. Under ‘Settings’, click ‘Advanced search’ and choose ‘Exact phrase’. Re-do the search and note the frequency figure given.
(ii) Repeat the ‘exact phrase’ searches 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/-ize spellings on the web.

(i) Go to Google, and type in realise normalise globalise in the search box. Click on ‘Settings’, click on ‘Advanced Search’ and choose ‘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 10.2 in the book. Can you see any patterns in the distribution of these spellings?