Introduction to using corpus linguistics

Corpus linguistics may be a completely new term to you. It may sound daunting but don't fret; this worksheet will help introduce you to corpus linguistics and how to go about using it. Language corpora (plural for corpus) can be useful tools to explore your topic of choice. What's more, in many cases, the data is already there for you, so you don't need to collect it yourself!

Section A: what is corpus linguistics?

Look up what is meant by 'corpus linguistics'. Write some notes and write a definition in your own words.

Where do I start?

- www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/introduction2.html – a general introduction.
- www.thoughtco.com/what-is-corpus-linguistics-1689936
- www.youtube.com/watch?v=32RjJ-lA-8Q – a detailed short animated video by Phloneme, a vlogger.
- www.youtube.com/watch?v=HQGcj2Cg-lY
  Prof. Tony McEnery from Lancaster University discusses corpus linguistics in a short video. (Lancaster have a free online course on the 'Future Learn' platform if you really want to get into corpus linguistics.)

What can I look at with corpus linguistics?

- **Word frequency** – how often is a word used in the corpus as a whole? Or in an individual text? Corpus software can give you list of word frequencies, with the most frequent words at the top.

- **Key word / Keyness** – a key word is a word that appears more frequently than you would statistically expect in everyday English. They can be useful in analysing a text. Corpus software can calculate the 'keyness' of words and then rank key words according to how statistically significant they are compared to everyday language use.

- **Key word in context (KWIC)** – how is the word/phrase you are looking at actually used? The KWIC shows you how the key word (or the word you are interested in) appears.

- **Collocation** – words that are often used together. Collocation can be measured statistically by a corpus and this can reveal attitudes (sometimes unnoticed) from the text producer.

- **Colligation** – Colligation means words are often used in particular grammatical structures. For example, the verb 'eat' in all its forms (eating, eats, ate etc.) tends to be used with a noun phrase as object ('the food that is being eaten').

Section B: using the British National Corpus (BNC)

What is the BNC?

The BNC is a general corpus that includes a wide variety of both written and spoken English in the UK (1980s-1993). A general corpus is usually extremely large with millions of 'tokens' (words and word forms) as it needs to represent everyday English.

You can access the BNC (and many other corpora) for free through the Lancaster University corpus query processor interface: cqpweb.lancs.ac.uk/ . It's free to register and gives you access to lots of different corpora. You can also find lots of video tutorials on YouTube: www.youtube.com/watch?v=0pSkr11xc1k&list=PL2XtJIhhrHNQgf4Dp6sckGZRU4NiUVw1e.

Once you are logged in, you can find lots of corpora, including the BNC. Click on the link to the BNC and you are ready to use it.

Make notes and write down your answers as you work your way through the following activities.

1.

a. On the left side of the screen is the query menu. Click on 'frequency lists' in the menu and leave the options on the next menu screen blank. Click on the grey button at the bottom to start the search.

b. What are the most frequent words in the BNC?

2. According to Zipf's law (a statistical pattern that was discovered in the 1930s by linguist George Zipf), the most frequent word is roughly twice as frequent as the second most frequent word, three times as frequent as the third most frequent word and so on. Does this apply to the BNC frequency list?

3.

a. With the BNC, you can just look at spoken language by selecting the 'restrictions' button in the standard search menu. This allows you to select the whole corpus - written or spoken data only. By clicking on 'restricted search' in the menu on the left, you can make very precise searches. For example, you can select English from a particular period in time or spoken language used by teenagers.

b. What selection would you need to make if you wished to investigate the spoken language of classroom interactions?

c. What selection(s) would you need to make if you wished to investigate the language of professions with regard to investigating how power affects interactions?

4.

a. With a partner, discuss the denotation (literal meaning) of 'juvenile' and then explore/list its connotations (implied meanings).

b. Now do a search for 'juvenile'. You will get a 'KWIC display', which may give you some sense of how the word 'juvenile' is used.

c. Next, click on the pull down menu in the top right hand corner and select 'collocations' as a new query. This will give you the statistically most significant words that 'juvenile' co-occurs with in the BN. Click on 'Go!' and you will get a screen that allows you to set the parameters including the type of statistical calculations you could use – just keep the settings as they are and click on 'create collocation database' at the bottom. What are the top five most frequent collocates for 'juvenile'?

d. How does this finding help you explain the difference you found between its denotation and connotations?

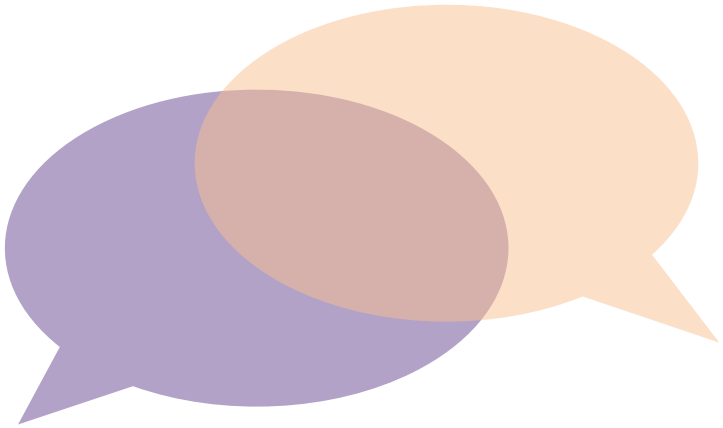
e. In the results screen, in the middle in the grey band with the bold heading 'Extra information' it tells you that the software used 'log-likelihood' (a type of statistical calculation) to calculate the collocation score. Look at the log-likelihood scores (column on far right) for the top five collocations for 'juvenile' – how do these figures help to explain the connotations of 'juvenile'?

5.

a. Using the 'standard query' function, select 'written data' in the 'restrictions' drop-down menu first. Then search for 'bloody'. Write down the frequency per million words (listed at the top of the results page) for this word in written texts.

b. Now do the same for 'bloody' in spoken data only. What do you notice about frequency of 'bloody' for written and spoken language? Can you explain the difference?

c. The BNC also gives you the total number of times the word 'bloody' appears in the written and the spoken data respectively. Why do you think you were not asked to look at the total number of times the word was used, but at the frequency pattern instead?



Section C: making language corpora work for your investigation

There are two ways in which you can use corpus linguistics. You can either use existing general corpora (e.g. BNC) or make your own corpus.

1. **General corpora**: existing corpora, such as the BNC, contain a huge number of texts from all modes. These are great for looking at English in general. You can access these online through interfaces such as CQPweb.

2. **Specialist corpora - DIY corpora**: If you are interested in a particular type of text, you could collect your own examples and build your own corpus (collection of texts) for analysis. There is a range of software available to download (often for free) that you can use to analyse your own corpus.

For each of the A-level topics below, list what kind(s) of corpus you could use? Briefly explain how you could use the corpus. As an extension you could consider other research questions linked to the topic, and how this might affect your choice of corpus.

	Topic	What kind of corpus? (General or DIY? Both? What sorts of texts would go in the DIY corpus?)	How could you use the corpus/corpora?
1	**Language and gender:** do men and women use swear words differently in speech?		
2	**Language varieties:** dialect/non standard English use of 'was'/ 'were stood'		
3	**Language varieties:** youth sociolect – what words do teenage speakers use that are less used by older adults?		
4	**Language and representation:** how are EU citizens living in the UK represented in the media following Brexit?		
5	**Child language acquisition:** what are the most commonly used word classes in caregiver speech?		