Library Guide:
Corpus Linguistics Resources

What is a Corpus?

A corpus is a body of texts which is available in a machine readable form. They can contain books, newspapers, speeches, and transcriptions of old texts. An example is the British National Corpus which contains examples of authentic language used both in speech and writing. Corpus software are used to interrogate texts to answer questions about language. They provide valuable insights into spoken and written language; word, phrase, and collocate frequencies; language variation and change in different countries over time.


There are many freely available corpora, but you will need to register to use them.

An example of a freely available corpus software is CQP (Corpus Query Processor) Web Server which has been produced and made available to students and academics by the University of Lancaster.

You will need to create an account via https://cqpweb.lancs.ac.uk/ to use this service.

You cannot access all corpora, so you will need to View Corpus Permissions to see what you can access. Examples of corpora that might be useful include: Corpus of English Dialogues, Early English Books Online, Works of Dickens, Shakespeare First Folio, and corpora of American and British English.

How to use a Corpus

Select a corpus to search e.g. Corpus of English Dialogues

The Simple Query Language syntax provides some clues as to how you can interrogate the corpus to ensure that you don’t miss anything. It is really important to understand how a corpus has been constructed to enable you to search it properly, so do read the help pages.

Type in search term e.g. library in the standard query box.

N.B. Now try searching for librar?? and you will notice different results. Understanding how wildcards work is essential to getting the most out of the corpus.

Click on Start Query

You will see the results displayed in the concordance. Click on the search term to see your key word in context.
“The greatest is the Enjoyment of my Friends, and such Worthy Gentlemen as your Selves, and when I cannot have enough of that; I have a Library, good Horses and good Musick.”

If you click on the hyperlink to the text, you can find out more about the source in which the key word is located. In this example, it is *The Lancashire Witches* a comedy by Thomas Shadwell, published in 1682.

You are also able to see how many texts have been searched for incidences of the word.

**Query mode and Restrictions**

You have the option to search case-sensitive or case-insensitive words e.g. if you were interested in how often the place name Bath is mentioned as opposed to how often people take a bath, you might want to run a case-sensitive search for *Bath*. You will not eliminate all examples of bathing as the word bath may be at the start of a sentence, but you can reduce the hits considerably. Similarly, if you were interested in whether it is usual to start a sentence with the word ‘However’, you could run a case sensitive search for *However* and use the *Restriction* menu to search only the *Written* part of the corpus. Other corpora allow you to restrict searches to broad genres, such as Fiction, Press, General Prose and Learned (Academic) texts, depending on the context you are interested in.

**Other Web-based Corpora**

Brigham Young University has a collection of corpora and corpus-based resources which allow you to see detailed entries for the top 60,000 words in English, enter and analyze text, download corpus-based data, including word frequencies, for offline use.


You will need to create an account via [https://corpus.byu.edu/](https://corpus.byu.edu/) to use this service.

The Google Books Corpus is useful for searching the frequency of words e.g. a search for Wincheap reveals that there are five books available online from the 1810s that include the word Wincheap and that Wincheap has been mentioned most in books published during the 1960s.

**Where to go for further help**

The university library has several books on corpus linguistics which will give you examples of how corpora can be used in research.

Overview of the Wikipedia Corpus [https://www.youtube.com/watch?v=-kzaOgrsvLk&rel=0](https://www.youtube.com/watch?v=-kzaOgrsvLk&rel=0)

CQPweb Tutorial [https://www.youtube.com/playlist?list=PL2XtJlhhrHNQgf4Dp6sckGZRU4NiUVw1e](https://www.youtube.com/playlist?list=PL2XtJlhhrHNQgf4Dp6sckGZRU4NiUVw1e)

Your Learning and Research Librarian will be able to help you make the best use of online resources.

E-mail learning.research.support@canterbury.ac.uk to arrange a convenient time to meet.