Using corpus to learn English (or any other language), by SSB

The aim of this post/talk/guide is to give you some information about the use of corpora (this is the plural form, the singular is *corpus*, Latin words are like this ;)) to improve your already fantastic English. So, I'll try to be as clear and useful as possible, and no too boring. Let's start.

What is a corpus?

Let's use dictionaries (by the way, the Oxford English Dictionary is my favourite, I'll refer to it as the OED):
Corpus: 1. a **collection of written texts**, especially the entire works of a particular author or a body of writing on a particular subject, AND ALSO 2. **a collection of written or spoken material in machine-readable form**, assembled for the purpose of linguistic research.

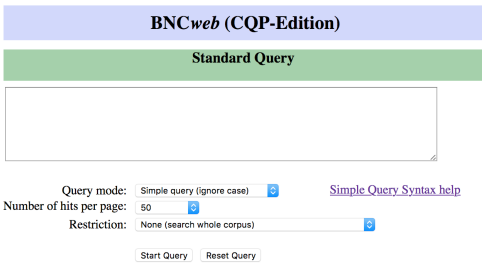
For our purpose, the second definition is more appropriate. Then, a **corpus** contains different texts (and when I say *text* this includes written and also spoken material)[1] which are arranged in one or several files (normally in txt format), and which can be used to search particular words or constructions that we are interested in.

So, **corpora** are compiled (that's the technical word, which means *created*) by gathering different texts, sometimes texts belonging to the same genre, for instance a corpus of novels published in the 20th century. On other occasions, **corpora** are built from different text-types, for example a bit of news published in newspapers, a bit of news published on the internet, a bit of news broadcasted on TV, and so on. All that would be a **corpus** of news.

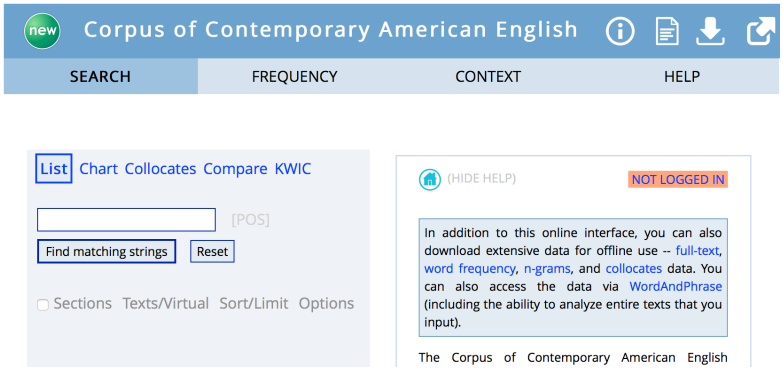
Corpora have been created with the main purpose of studying the language. It is like the forest for the study of how different species interact among them, or for discovering new species. Thus, by means of **corpora** linguists can research how language changes with time, which new words are created, new combinations of words, pronunciation patterns, differences among varieties of the same language, e.g. US American vs. British English, but also Australian vs. South African English, and many more things.

Just two more features: 1. **corpora** are available as txt files to be used on a computer with a special software (but also with a normal text editor program), and some of them can also be used online. 2. There are synchronic **corpora**, i.e. those which include texts belonging to more or less the same period of time (20 years), and diachronic or historical **corpora** that are compiled with material from a broader range of time, i.e. two centuries.
Let's illustrate this a bit:

1. British National Corpus (BNC)
2. Corpus of Contemporary American English (COCA)



[1] Spoken texts are collected by recording people's talk and transcribing the record, i.e. writing down the spoken words. Some transcriptions even include features such as intonation, stops, and other interesting features of the spoken language.



What can we do with corpora?

Simply, we can learn/know things about English, for instance:

1. collocation patterns: this word is normally used with this word, but NOT with this other. e.g. which verb collocates more frequently with the noun "information". This is the result obtained from BNC.

Frequency breakdown of lexical items for position "node" (789 types and 4249 tokens)

|< << >> >| Frequency breakdown of tags only Go! Download whole table

No.	Lexical items	No. of occurrences	Percent
1	provide information	340	8%
2	providing information	106	2.49%
3	give information	103	2.42%
4	provides information	76	1.79%
5	have information	71	1.67%
6	obtain information	71	1.67%
7	get information	63	1.48%
8	gathering information	61	1.44%
9	is information	61	1.44%
10	giving information	59	1.39%
11	gather information	56	1.32%
12	collect information	55	1.29%
13	using information	55	1.29%
14	contains information	45	1.06%
15	contain information	43	1.01%
16	collecting information	42	0.99%
17	receive information	41	0.96%
18	need information	41	0.96%
19	share information	41	0.96%
20	include information	41	0.96%
21	exchange information	38	0.89%
22	obtaining information	34	0.8%

2. Which preposition is the most appropriate for one word?, e.g. build + prep

3. Which is the correct/appropriate word order?

She arrives always late vs. She always arrives late

4. New trendy expressions

These are some of the applications for learners of English, but there are many more. In fact, an important part of the content of textbooks for language learning uses material from **corpora.**

What types of corpora can we have access to?

As I mentioned before there is a great variety of **corpora**, some of them are very specialized, others are only available for academic purposes (you have to apply or even pay for them), but fortunately there are others which are available for free, ☺.

This is a list of some of them:

1. British National Corpus (BNC): compilation of written (60%) and spoken (40%) of British English, mostly from the 90s till the first decade of the 20th century. Free. 100 million words! 2 links: http://corpus.byu.edu/bnc/ or http://bncweb.lancs.ac.uk/cgi-binbncXML/BNCquery.pl?theQuery=search&urlTest=yes
2. Corpus of Contemporary American English (COCA): compilation of written (50%) and spoken (50%) of American English from 1995 to the 20th century. Free (you have to register and it gives a maximum of 50 searches per day). 522 million words!! http://corpus.byu.edu/coca/
3. Corpus of Global Web-Based English (Glowbe): text obtained from the internet belonging to 20 different varieties of English, 1.9 billion words. http://corpus.byu.edu/glowbe/
4. News on the Web (NOW): news from the web, started in 2005. It is updated every day. 2,5 Billion words. http://corpus.byu.edu/now/
5. International Corpus of English (ICE): Several "small" corpora of different varieties of English, e.g. from New Zealand, Australian, British, American, from Singapore, Indian, from Hong Kong, the Philippines, South African, Nigerian, and so on. They were mostly compiled in the 90s, some a bit later. The project is still developing. Each corpus is composed by 60% spoken and 40% written materials. I think most of them are free but you have to ask for them. Then they send you a link and you can download the corpus. It must be used in your computer.

How to use corpora

More or less all the online **corpora** work pretty much the same, so I'll show you the steps for the BNC and the COCA, and later you can explore yourself the rest of options. There is no way you can break anything, so don't be shy and click wherever you like. They usually include very good help files, but sometimes the language is a bit technical.

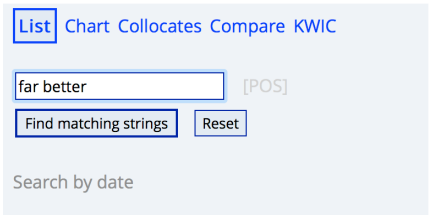
BNC guide

1. Go to the BNC
2. just type the word you want to see "in context" in the main box and click "start query"
3. Now you have the answer in KWIC format (Key Word in Context), that is, the word you have typed in the centre and the rest of the text at the left and the right side.

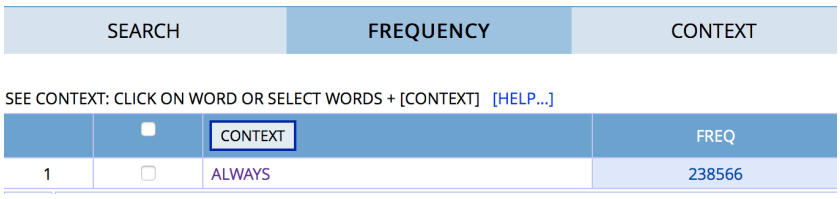
4. To order this information click on the upper-right window and select "sort". This takes you to another window where the "key" word is again in the centre but the sentences have been ordered according to the first word to the right side (alphabetically). At the top, you can select other ordering options. Imagine you want to know which words come before your key word, so you can order the result by the first word to the left.
5. To know more about how the key word is used in the corpus you can go to "collocation", in the same window as in step 4, then click "submit", to obtain the words that collocate with your key word at both sides. You can change the options in "collocation window span". **This is interesting for language learners because normally the most frequent option in a corpus is the more appropriate in a native context.**
6. Apart from this, it is possible to do more complex searches, for instance, by word class, e.g. all the verbs with the pattern verb+to+ *ing* form. In order to do this type of searches you need to know the code that each word class has been assigned, so you have to go to the help file. This is just an example:

COCA guide

1. The COCA has a very similar website. The principle is always the same. In this case, you have to sign up before in the upper-right windows, although a limited number of searches are permitted without logging.
2. Then go to "search" and type the word in the white box and click "find matching string". Tip: you can search for a word or for a string of them. See:



3. you obtain the result in the "frequency" label



4. If your click the word, the corpus shows you the word in context, in the KWIC style.