Tutorial for using online corpora

at corpus.byu.edu

The corpora to use

The website corpus.byu.edu offers access to several different linguistic corpora via the same interface. This interface has many useful features; however, you will only need to use a few of them. These features are explained here, with special reference to the tasks required for the three studies in Assignment 3.

To get to the corpora, open a web browser and go to http://corpus.byu.edu. Note: if you have trouble with one web browser, try another! For example, Firefox works a bit better than Internet Explorer. Then select the corpus you want to search. The three that are relevant for Assignment 3 are these:

- *Time* Magazine Corpus of American English
- Corpus of Contemporary American English (COCA)
- BYU-BNC: British National Corpus (NOTE: click only on the BYU-BNC part; otherwise, you are sent to the original BNC website.)

The TIME corpus shows the results broken down by decade (1920s, 1930s, etc.), while the BNC shows the results broken down by register (spoken, fiction, academic, etc.). COCA shows the results broken down both by time and by register. (But note that you must tick the SHOW box next to SECTIONS for any of this to happen.)

At first, searching any of these corpora is free. However, after you have done a few searches, you will be asked to create an account. You might as well go ahead and do this right away. It’s free, and it gives you access to all of the corpora (though with a limit of 100 searches per day). Then you should log in before you start searching.

Using the interface

Once you get to the main search interface, you will see that there are three main areas: on the left is the area where you specify your search terms. At the top right is the area where the results are shown in summary form, and at the bottom right is the area where the individual hits are shown. To see hits here, you need to click on one of the links in the results summary area.

Note that every little section of the page has little “?” icons; these link to help pages that explain how to use the interface. These supply all the information you need, but read on for more targeted instructions.

Type your search string (i.e., the word you’re looking for) in the WORD(S) box under SEARCH STRING. You can perform a simple search by typing in a word here and clicking SEARCH (or hitting the Enter key).

However, for this assignment (whichever option you choose), you will need a couple more settings. For one thing, because the corpora (and corpus sections) are different sizes, you will need to normalize your results—that is, you don’t want the total hits, but rather the hits per million words, which is a conventional measure in corpus linguistics. (See the separate handout on normalization.) To have the website normalize the results for you, click on the text at bottom left that says CLICK TO SEE OPTIONS. Four more lines will appear. In the DISPLAY box, select PER MIL, which means “per million words”. You should do this every time you start a session. Note that normalized results will always involve a decimal point—if your results don’t, they may be wrong!

Let’s look at a couple of interesting features of the interface. Start by going to the BYU-BNC, as explained above. First, set the DISPLAY option to PER MIL, and then make sure the box called SHOW next to the word SECTIONS is checked. Then type dog into the WORD(S) box, and click SEARCH. You should get some results at top right, showing that the total number of hits is 7846 (meaning that “dog” occurs that many times in this 100-million-word corpus). It also breaks the results down by section, showing, for
example, that in the spoken part of the corpus, “dog” occurs 133.08 times per million words (note the two decimal places). You can see that dog is therefore over four times as frequent in the spoken part of the corpus than in the academic writing part, where it has a frequency of only 28.37 per million words. If you see different colors in the background, these indicate higher and lower values—just an aid in recognizing patterns.

Using wildcards

Now perform the same search, but with the word dog*—that is, with an asterisk (or “star”, for the uninitiated) at the end. What you should see in the results area is more complex; this time, we see all the words that begin with “dog” in the corpus. The top row, for plain old “dog”, should show the same results as last time. The asterisk is called a wildcard and stands for “whatever”. Try the following search: dog * dog (with spaces around the asterisk). This time, the asterisk stands for a whole word, yielding as the top hit “dog eat dog”, as you might expect.

Note also the existence on the list of “dog”, dog” and similar hits. This is because in a corpus, the punctuation is separated from the words, to make searches more accurate. This is called tokenizing the corpus. This means that if you want to search for a punctuation mark such as a comma or full stop, you should leave spaces around it. Importantly, this also means that if you want to search for an s-possessive (such as dog’s), you will need to put a space before the apostrophe (as in dog’s).

For more information on the query language, click the “?” symbol to the right of the WORD(S) box and read the page that opens.

It is also possible to limit the search to words that have a particular part of speech (or word class), which is abbreviated POS. This is made easier by the POS LIST below the WORD(S) box. Let’s look for the combination ADJ + dog. Put the cursor in the WORD(S) box by clicking in there, then click on POS LIST to show the list, and select adj.ALL in the list. The text [aj*] should appear in the WORD(S) box. (Note that all you really have to do is type this, once you have learned the codes.) After this text, type a space and then dog, so that you have [aj*] dog. Click SEARCH. You should get a list of results in which little dog is the most common, and hot dog is the eighth most common.

Now let’s search for the verb dog only. In the WORD(S) box, type dog.[v*]—with no space on either side of the period (dot). This time you are not searching for dog and a verb, you’re searching for the verb dog. You should get only 22 hits, which is not very many in a corpus of 100 million words. To see the individual hits, click on the word dog in the results list. They should appear below. If you look through these, you will see that many of them were misclassified as verbs. This should serve as a warning—corpus tools have their limits! The automatic part-of-speech tagger used with this corpus is very good but can be thrown off by certain types of contexts.

Getting a random selection

There is one last thing you need to learn before you will be prepared to do Assignment 3. That is how to get a random sample of hits. Search again for plain old dog; you should get 7846 hits again. Let’s say we want to look at 20 of these, but a random sample, not the first 20. Click on the word dog in the hits, so that the first page of examples appears below. There are 100 of these, but note that they are all from tabloid newspapers—not what we want. Look at the top right portion of the examples area, where it says SAMPLE: 100 200 500 1000. Click on the number 100. This will give you another set of 100 examples, but from all different parts of the corpus (if you had originally searched a sub-part of the corpus, they would be randomly drawn from only that part.) Since these are randomly drawn from the corpus and randomly ordered, you can take the first 20 hits in this list and be assured of getting a random sample.

Comparing corpora

One last tip that makes using the BYU interface especially attractive: You can run the same search on a different corpus with one click (at least in some browsers)! Look at the top right of the results summary area, where it says COMPARE: COCA COHA TIME BNC GOOGLE. By clicking one of these corpus names, you can have the website run the same search in the selected corpus. Very useful if you are doing Topic B!

Good luck, and have fun! Hopefully, you will come to realize how useful corpora can be for answering your own questions about English...