Methods for Text Mining and Analysis of Text Corpora

Basic Text Analysis – Part II

*Some of the exercises here and in the rest of the workshop will use the tool AntConc, which is a powerful and simple tool to begin text analysis. This, along with other files which will be needed, is available in a folder for download, or can be accessed online. Some text files to experiment with are also provided, including the Brown corpus.*

***Word frequencies***

Start AntConc and use the ‘File > Open file(s)’ menu option to load one of the sample texts. Then select this in the files window on the left hand side of the tool, and use the ‘File view’ function to make sure that it is properly loaded. You will have to click the ‘Start’ button.

Change to the ‘Word list’ function and click start. Make sure that the ‘Sort by’ option at the foot of the screen is set to ‘Sort by Freq’ (i.e. frequency, the number of times the word appears).

The tool will list the most frequent words in your text along with a count of the number of times they appear. Either write down the top ten results, or click the ‘Clone Results’ button in the bottom right-hand corner.

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

Now clear this text and load a different one. Perform the same routine (remember that you have to click ‘Start’ each time you return to an option you’ve used before). Compare the ten most frequent words in this text to that from your previous text. Which ones are the same?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

*The purpose of this exercise is to demonstrate that in any text written in English, the most-used words will almost always be very similar. Most researchers will not be interested in these words (although there are times they will be), and so there needs to be a way of removing them.*

***Stop Lists***

Locate the file called ‘Basic Stoplist.txt’. In AntConc, go to the ‘Settings’ menu, select ‘Tool Preferences’, and then ‘Word List’ in the window that appears.

Under the ‘Word List Range’ options, select ‘Use a stoplist below’. Click ‘Open’ next to the ‘Add Words from File’ option and navigate to the ‘Basic Stoplist.txt’ file.

Once the wordlist has loaded in the window, click ‘Apply’ at the bottom of the window.

Return to the ‘Word List’ function in AntConc and click ‘Start’ to get it to create a new version of the wordlist. Have a look at the results. Are there any other words that you might add to the ‘Basic Stoplist.txt’ file?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

Think about the way in which the words that are in the stoplist function in texts. Can you think of any occasions on which you might want some of these words NOT to be stoplisted?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

***Sorting Concordance Lines***

Using AntConc, go to the ‘Concordance’ tool (N.B. If you start from this point, you will have had to load a file and create a word list based on it).

Choose a word to enter into the ‘Search Term’ box at the foot of the window. This can be any word you like. You might want to use one of the most frequent words which resulted from the previous exercise (although ideally not one of the words that would be stoplisted, unless you especially want to!). If you’re feeling very unimaginative, try *man* or *woman* – or one and then the other to compare them!

The tool will return a list of all the times your search term appears in the text you’re looking at. The search term will appear in blue in the centre of the line, and a stretch of text will be provided to either side to give context.

Choose one of the lines that are returned and click on your blue search term in that line. AntConc should jump to the ‘File View’ tool and show you where this particular instance appears in the text. To return to the concordance lines, click back on the ‘Concordance’ tab at the top of the AntConc window.

The results you have seen so far are returned in the order in which they appear in the text. Next sort the results by clicking on the ‘Sort’ button underneath the search term box. The default settings will look like this



Under the ‘Kwic Sort’ options, the different ‘Levels’ list the words on which the ordering is sorted. The default means they are sorted by the first word to the right of your search word, then the second word to the right, then the third word to the right. You can deselect the ‘Level 2’ and ‘Level 3’ boxes if you wish by clicking on the blue tick boxes.

Skim through your results. Is there anything interesting about the words which appear immediately after your search term? Might they tell you anything about how that word is being used in this text? If your search term is a noun, how often is it followed by *is* and what words appear after this? Don’t worry about spending too much time on this, as we will return to these questions later in the workshop.

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

Next, deselect ‘Level 2’ and ‘Level 3’. Then change the contents of the ‘Level 1’ box to ‘1L’ – you can do this by using the down arrows next to the box. Then click ‘Sort’ again.

Once you have these results, have a look at the words immediately to the left of your search word. Do these tell you anything about the use of the word? If your search word is a noun (such as *man*, *woman*, *table*, *mountain*), look especially for adjectives which precede the word.

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

*These exercises are the beginnings of looking at ‘collocates’. Collocates are words which appear with a word that you are interested in. By ordering the words on the 1R or 1L positions, you are effectively looking at the collocates of that word in those specific positions. This gets more complicated once you go further out from these positions, and to do that, AntConc uses a different tool*.

***Collocates***

Using AntConc, select the ‘Collocate’ tool from the list of options at the top of the window. In the ‘Search Term’ box, enter one of the words which you searched for in the previous exercise.

Look for the ‘Min. Collocate Frequency’ option in the bottom right of the window. Change this to ‘5’ initially. This means that collocates which appear fewer than five times will not be listed in the results. Leave the ‘Sort by’ option as ‘Sort by Stat’ – you will learn more about what these mean in the next session.

Click ‘Start’. The collocates that are returned are words which appear (by default) in a stretch of words beginning at five words to the left of your search term and ending five words to the right of your search term. Have a look through these. As before, do they suggest anything to you about the way your search term is used in the text, and therefore potentially how it is viewed by the author(s) of your text?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

Next, it is interesting to adjust the span of text around your search term in which you ask the computer to find those collocates. Look at the ‘Window Span’ options, which adjust this. Try changing the values to ‘From 1L to 1R’ and see what difference this makes to your results. Then try going in the opposite direction and changing the values to ‘From 10L to 10R’ and seeing what this results in again. Remember that you can click the ‘Clone Results’ button at any point to help you make these comparisons by copying your results into a new window.

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

………………………………………………………………………………………………………………………………………………

*There is a degree of debate in how large a window span can be used and still return meaningful results. The ‘Stats’ help you to decide how important a result you obtain is, but at the same time there is a general consensus amongst researchers that beyond five words to either side of your search term, the results become progressively less reliable.*

***Collocates in Online Corpora***

Keep AntConc open with your text still loaded.

Go to the website corpus.byu.edu/bnc

Remember that you will need to register with the BYU Corpora website to be able to carry out more than a few searches using the site.

Type a search term into the search box on the page that appears for the BNC (British National Corpus). You can use one of the words that you used in the previous exercises, or you can choose a new word that you think might be interesting. It is worth trying these exercises with different types of word such as nouns (e.g. *woman*), adjectives (e.g. *black*), and verbs (e.g. *criticise*).

If you want to, try running basic searches for these words and looking at some of the results in the concordance lines before we move to the stage of looking at collocates.

On the search page, select the ‘Collocates’ option. Type your search term into the ‘Word/phrase’ box and, to begin with, put an \* into the ‘Collocates’ box. Look at the list of collocates which are returned.

If you still have AntConc open with the previous text, go back to it and change the ‘Window Span’ to ‘From 4L to 4R’ (as this is the maximum the BNC interface will allow). Search for the same search term and its collocates in both your AntConc text and the BNC. What are the top five collocates for the same term in each of the two different text collections? Are there any that are the same?

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

……………………………………………………………………………………………………………………………………………….

*Collocates can tell us a lot about the way in which a word is viewed by the writer or speakers who are using it. Large corpora such as the British National Corpus can be used to get a broader view of the way a word is used in society – although it is always important to remember that the limitations to representativeness of data always mean that it is difficult to say that a society, a culture, or even an individual writer ‘thinks’ something unequivocally.*