Corpus Linguistics with AntConc // Heather Froehlich // @heatherfro

0. Aims and goals: by the end of this tutorial you will be able to

- create/download a corpus of texts
- conduct a keyword-in-context search
- identify patterns surrounding a particular word
- use more specific search queries
- look at statistically significant differences between corpora
- make multi-modal comparisons using corpus lingiustic methods

1. Using concordance software

AntConc is one of several concordance software programs. Like Chrome vs Firefox or iPhone vs Android, they each have their strengths and everyone has their own preferences. The principles discussed today extend beyond Antconc specifically, though the how-to details may be slightly different in different software packages.

- AntConc works only with plain-text files with the file appendix **.txt**
- It will read XML files saved as .txt files
- Concordance software will take every character in a text file into account – it may be necessary to clear out repeated "boilerplate" information or metadata, as it may skew your results!

2. Getting familiar with the user interface

There are 7 tabs in the AntConc user interface, plus a window to see loaded corpus files.

- **Concordance:** This will show you what's known as a **K**ey**w**ord **i**n **C**ontext view (abbreviated KWIC), using the search bar below it.

- **Concordance Plot:** This will show you a very simple visualization of your KWIC search, where each instance will be represented as a little black line from beginning to end of each file containing the search term.

- **File View:** This will show you a full file view for larger context of a result.

- **Clusters:** This view shows you words which very frequently appear together.

- **Collocates:** Clusters show us words which *definitely* appear together in a corpus; collocates show words which are *statistically* likely to appear together.

- **Word list:** All the words in your corpus.

- **Keyword List:** This will show comparisons between two corpora

2. Searching

Antconc operates character searches, which means it takes every keystroke into account, and can be sorted alphabetically to the left and to the right.

Some special characters allow for more robust searches:

*	Zero or more characters
+	Zero or any one character
?	Any one character
@	Zero or one word
#	Any one word
|	Search term ‘OR’ search term

The ? operator is more specific than the * operator:
wom?n = both women and woman
m?n = man and men, but also min
contrast to **m*n:** not helpful, because you’ll also get mean, melon, etc in your search!

Save your results as **File > Save output to text file** (& append with .txt). They can then be opened as a plain text file.

3. More advanced analyses

3.1 Collocation: what words are most likely to appear near a specific search term?
Operates on conditional probability, broadly expressed as
P(w1, w2) and P(w2, w1)
2nd word in bigram appears given the 1st word
1st word in bigram appears given the 2nd word, and computed on a scale of 0-1.
AntConc uses a measure called Mutual Information Scores to compute this.

3.2 Comparing corpora: your corpus against a suitable reference corpus
(different reference corpora pull out different kinds of features in comparison!)
Settings > Tool preferences > Keyword List
Under 'Reference Corpus' make sure "**Use raw files**" is checked
Add Directory > open the folder containing the files that make up the reference corpus.
Hit **Load** (& wait …) then once the 'Loaded' box is checked, hit **Apply.** (Make sure you have a whole list of files!)

You can also opt to swap reference corpus & main files (SWAP REF/MAIN FILES) and it is worth looking at what both results show.

In Keyword List, just hit Start (with nothing typed in the search box). We see a list of Keywords that have words that are much more "unusual" – more statistically unexpected – in the corpus we are looking at when compared to the reference corpus. These terms are measured by **keyness.**

Keyness: this is the frequency of a word in the text when compared with its frequency in a reference corpus, "such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a p value specified by the user” --
http://www.lexically.net/downloads/version6/HTML/index.html?keyness_definition.htm

4. An Introductory Bibliography to Corpus Linguistics:

http://hfroehli.ch/2014/05/11/intro-bibliography-corpus-linguistics/