Using Antconc and Sketch Engine in your own language investigation

If you wish to use both or one of these tools in your own investigation, here is a useful summary of what each tool allows you to do, as well as some further suggestions and ideas for using Antconc and Sketch Engine with your own data.

Using corpus tools is very similar to using a calculator in maths or sciences. The technology does not do the critical thinking and the analysis for you. It is still down to you to look at your data and ask sensible questions.

Comparison of Antconc and Sketch Engine

All concordancing software and interfaces allow you to investigate the same basic characteristics of a corpus. Below is a reminder of the basic corpus functions in Antconc and Sketch Engine.

Functions that both **Antconc** and **Sketch Engine** have and have similar ease of use	
1.	**Concordance, KWIC (key word in context)** Both of these corpus tools allow you to produce concordances, where you get a display with the KWIC, which is often helpful with regard to exploring issues in a corpus. For example, if you want to explore what the attitudes are towards immigrants in a corpus of newspaper articles that you've collected, you can search for 'immigrant' and see almost immediately if there are particular collocations or other patterns. This can then guide your further investigation of the data.
2.	**Frequency list, collocates** Both allow you to create a frequency list. Most of the words on a frequency list will be function words but any unusual words such adjectives, proper nouns or lexical verbs appearing in a frequency list are good places for further investigation. Collocates give you a measure of how strong the co-occurrence of your search term is with other words (collocations) in the corpus. Recent research with a corpus of British news articles on immigration found that the noun 'immigrant' had a very strong collocation pattern with the attributive adjective 'illegal' in the British press. 'Illegal' was the most common collocate for 'immigrant' in most of the newspapers. This gives you a good idea about the attitude towards immigrants that the British press is conveying.

However, different corpus tools do also differ in terms of additional functions that they offer. Here is a list of features that are typical for Antconc and Sketch Engine. With these functions, you can use both Antconc and Sketch Engine with your data or your own corpus.

Antconc only functions	**Sketch Engine** only functions
Concordance plot This gives you a visual representation of where the word/token you are looking for is used across the whole corpus. Sometimes a word only appears in a few locations, is very much spread evenly through a corpus or is very dense in only a few places.	**Key word / keyness** (This is also possible with Antconc but only if you have a reference corpus yourself.) If you want to see if your data/corpus contains word(s) which appear more frequently than is statistically likely based on everyday English usage, then you need to use Sketch Engine to calculate 'keyness', because as an online interface it gives you access to several potential reference corpora, for example the BNC and Web English 2013.

Antconc only functions	**Sketch Engine** only functions
Collocates Of course you can search for collocates with Sketch Engine, but Antconc is more user friendly in this area. You can actually specify how many words to the left and to the right of the node word (the word that you are investigating) you want your search to include. Often collocates do not consist of a neat side by side pairing of adjective-noun, but usually five places to the left and five places to the right of the node is a good measure to find collocates.	**Word sketch** As all corpora in Sketch Engine are automatically tagged for 'part of speech' (each word in the corpus is labelled with a word class and syntactic function). The program 'reads' the tags and uses them to create word sketches that show you how a word is used in the corpus. Is it used to modify nouns? Is it the subject or object of verbs? A word sketch allows you to also investigate the grammatical aspects of the word(s) you are interested in.
Clusters / N-grams Sketch Engine also has this but Antconc is much easier to use. A cluster or N-gram is a special form of collocation. It is useful to see if a word you are interested in is part of a set phrase (the cluster or N-gram). For example, looking at data from *Star Wars*, it is clear that the noun 'jedi' is going to be significant. A concordance of 'jedi' will give you a list of every time this noun is used in the scripts, but it is hard-going to check by hand if it is part of a longer, set phrase. So by switching to 'Cluster'/'N-gram' at the top of the toolbar task tabs, you immediately get any set phrases featuring 'jedi' and it turns out that it forms part of the title 'jedi knight' typically. Antconc has calculated that 'jedi knight' is a set phrase, not just a collocate pairing, which is more accidental than a N-gram. 'Jedi knight' isn't simply a collocation. It's a lexeme carrying a distinctive meaning. This is similar to phrasal verbs - 'give' means something completely different from 'give up', so 'give up' isn't just a collocation, it is a N-gram, a set phrase which is actually a distinctive word with its own meaning.	**Access to corpora** As you can see in the home page menu, there is a very long list of corpora. This means that you do not have to collect texts and make your own corpus. You may find that there are interesting corpora already uploaded. Child language or newspaper data are among the examples that may be of interest for your investigation. If you are not sure what a corpus is about or what type of texts it contains, you can click on the 'i' symbol to get the details about the corpus.
	WebBootCat This is a tool that allows you to search for words or phrases present in texts on the internet. The tool will 'harvest' any online text with the search term and you can quickly build your own corpus this way. The building of the corpus includes tagging the texts (so you can create word sketches etc).

DIY corpus tools for language investigation

1. British National Corpus (via CQPWeb)

Even if you are not interested in using corpus tools such as Antconc and Sketch Engine, you can still use the BNC corpus via CQPWeb.

Imagine that you have made transcripts of interactions between sixth form students. You can analyse these using the language levels and linguistic concepts – however, it may be that you find a particular noun or adjective is used by speakers in a non standard fashion. You can test this by looking up the word in question in the BNC. By getting a concordance of the word you are interested in you will find whether the speakers in your data are indeed using it differently from everyday English.

For example, say that the teenage speakers in your data use the adjective 'wicked' as a positive evaluation – a look at the collocates for 'wicked' in the BNC shows that while it's used positively, 'wicked's' principal meaning/usage is still synonymous with 'bad' or 'evil'. This suggests that the use of 'wicked' as a positive evaluative adjective is distinctive for the teenage speakers in your data. This usage is part of their sociolect.

2. Putting corpus tools at the heart of your methodology

If you wish, you can make using corpus tools a central part of your investigation. Collect texts from the sources or on the topic that you wish to investigate. As the language investigation is fairly small, keep your data set / corpus manageable.

Although you can use Sketch Engine's 'WebBootCat' tool to harvest texts from the internet to investigate, you may wish to use smaller data sets so that you can still analyse/annotate your data in order to explore other language features.

Corpus tools are very helpful in lexical/semantic analysis and in exploring whole discourses (corpus tools are also successfully used to explore grammatical structures, but this is more challenging – although you can do some of this with the 'word sketch' function in Sketch Engine).

Some types of data that are well suited to corpus analysis:

- Political speeches (easy to find online and downloadable).
- Emails, blogs, comments on YouTube videos, social media messages. For example, you could look at tweets using the hashtag option as a means of selecting particular data.
- Extracts from fiction or even whole novels. Project Gutenberg has digital formats of classic texts. You could collect some or all the novels of a particular writer e.g. Charles Dickens.
- Song lyrics. If you wanted to explore the attitudes expressed in a particular musical genre, you could collect all the song lyrics from the most representative artists in the genre or you could focus on a particular artist e.g. Beyoncé's lyrics to explore her representation of women.
- Film scripts. For example a particular genre or featuring the same actor as the central character etc.
- Recent newspaper articles on a topic that interests you or that is divisive. Both the Guardian and The Daily Mail allow you to search for topics and you could select the 15 most recent articles on your choice of topic. These two papers are actually quite interesting to compare as they are from opposite ends of the political spectrum.

3. Formatting your data

Make sure to save your texts in a format that can be read by the software. Plain text files work with any concordancing program. You may also wish to keep the original formatting, as these corpus tools don't allow you to look at the graphological features and text image cohesion, which affects meanings.

4. Getting started

Choose a program to start your analysis. Create a frequency list first to see if there are any lexical words among the function words. These would make good words to create concordances and collocates for.

5. Concordance plot in Antconc

Words that are central to your data or corpus are also useful to look at in the 'concordance plot' function of Antconc. Do the word(s) only appear in a few places in the corpus or is it more widespread? If your corpus consists of separate texts (e.g. all the newspaper articles from one week about a news story), Antconc will give you a concordance plot for each separate text. Again, do the word(s) appear only in certain texts and not in others? Why might this be?

6. Keyness in Sketch Engine

Key words are useful for any corpus investigation. So, once you have your files/corpus, upload them to Sketch Engine as a corpus and execute a keyword calculation. The keywords will give you a real sense as to what the data is really about or concerned with. E.g. a keyword analysis of Star Wars; A New Hope shows that 'Luke' is the number one keyword for the script. This suggests that the story of the film is mostly the story of one character.

7. Word sketches

It is very useful to explore in what grammatical constructions important or frequent words are used in the corpus. For example, when looking at news reports about crime, there may be lots of proper nouns such as the names of the victims, which suggests that the victims are given some importance. However, a word sketch of a victim's name may show you that it is frequently used in passive constructions, which suggests that in the media the victim is represented as helpless.

Useful links for video tutorials on Antconc and Sketch Engine

There are lots of support materials available to help you get started with using corpus tools. A quick search will usually take you to several universities' linguistic departments' sites, which often have free downloadable user guides for working with corpus tools.

Some useful links to tutorials

1. Laurence Anthony's YouTube Channel, called 'AntLab', where he has posted video tutorials on working with Antconc: www.youtube.com/user/AntlabJPN

2. The team behind Sketch Engine has also got its own YouTube Channel, where you will find video tutorials: www.youtube.com/channel/UCw_-94XlIh-a2-i2GrgjvEQ