Methods for Text Mining and Analysis of Text Corpora

Stylistics

***Semantic Prosody***

*Much of the material we have looked at yesterday has helped us to get an impression of the associations which exist between words in text. This can be very useful for telling you about attitudes towards some of these ideas, including some which may not be automatically apparent to you.*

Go to the BNC on the BYU website – corpus.byu.edu/bnc – and type ‘utter\_jj’ into the search box. The results you get will be ordered by the text in which they appear, and a more random sample is needed. At the top of the results page, click ‘100’ next to the ‘FIND SAMPLE’ option.

Look through the first twenty concordance lines which are returned to you, and for each find the first noun which follows *utter* and write these down:

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

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

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

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

What does this tell you about what people associate with *utter*? Is it generally used for positive or negative situations?

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

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

Look at the following Philip Larkin poem, which has been discussed in the light of similar findings to those above:

Lambs that learn to walk in snow

When their bleating clouds the air

Meet a vast unwelcome, know

Nothing but a sunless glare.

Newly stumbling to and fro

All they find, outside the fold,

Is a wretched width of cold.

As they wait beside the ewe,

Her fleeces wetly caked, there lies

Hidden round them, waiting too,

Earth’s immeasurable surprise.

They could not grasp it if they knew,

What so soon will wake and grow

Utterly unlike the snow.

 *First Sight*, Philip Larkin

Specific attention has been paid to the word *utterly* in the final line. What do you think the poet is talking about? What is it, on the surface, which appears to be ‘utterly unlike the snow’? Do you feel, instinctively that this is a positive or a negative poem?

Search for *utterly* in the BNC. Does it seem to be positive or negative? Does this either reinforce or change your feelings about whether the poem as a whole is positive or negative?

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

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

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

*These patterns provide what is known as the ‘semantic prosody’ of a word – a feeling of the emotional weight attached to it. A slightly different aspect of this is what is known as ‘semantic preference’, which is related to patterns of ideas which are related to a word. This is a development of the collocate searching which we were looking at yesterday, but is interested in broad patterns of association*.

To see semantic preference in action, try searching for the phrase *naked eye* in the BNC. What types of word seem to be associated with this phrase? Can you separate them into any types of group? What would you consider to be something that might break this pattern?

*In linguistic analyses of literature, usually known as ‘Stylistics’, breaking a pattern may be evidence of an author ‘foregrounding’ some aspect of what they are saying. This means that there is a value both to knowing what is ‘normal’ and what is ‘not normal’ for speakers of a language to do.*

Try similar searches with the following verbs and see if you notice anything interesting:

*commit*, *brook*, *budge*, *incur*

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

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

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

Next, return to AntConc and load some of the text(s) you have been working with, either ones that you have brought with you, or ones that you have selected from the Project Gutenberg repository.

Use the ‘Keywords’ tool to find some of the key items in your texts, using the Brown corpus as your reference.

Switch back to the BNC and search for your key terms. Can you identify anything about the semantic prosody or the semantic preference for these words from the BNC interface?

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

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

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

Now look at the concordance lines for these keywords returned by AntConc from your text. To what extent does the use of these keywords by your author appear to agree with their general use as indicated by the BNC? Does this tell you anything about the way your author is representing different terms, or does it indicate that their views of these terms may differ from more general societal trends as indicated by the BNC?

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

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

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

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

 ***Characterisation***

*The Corpus Linguistics in Context (CLiC) project has built a free resource and interface which allows users to investigate corpora of Victorian novels. They have two main collections, one of all of the novels of Charles Dickens, and one of 19th Century children’s literature. However, they also have a number of other Victorian novels which can either be investigated in their own right or used as part of a ‘reference’ corpus.*

*One of the primary interests of the researchers creating this resource was the way in which ‘suspensions’ can be used by authors to add characterization details to people speaking. A suspension is an interruption to direct speech, and often includes detail of something the speaker does or of their emotional state.*

Go to the CLiC application’s website: clic.bham.ac.uk/concordance

The interface takes a reasonable amount of time to become fully familiar with, and we strongly encourage you to return to it later and read other material provided on the page to get an idea of how to use it more extensively.

In the menu on the right-hand side of the page, the ‘Concordance’ option should be automatically selected, but scroll down to see what the other options are.

In the ‘Concordance’ option, set ‘Search the corpora’ to ‘DNov – Dickens’s Novels’; set ‘Only in subsets’ to ‘Long suspensions’; and in the ‘Search for terms’ box enter ‘hand’.

This should now show you all of the instances from Charles Dickens novels in which the word *hand* appears in a long suspension (i.e. one that is more than three words long). Have a look through the results to get an idea of what characters do with their hands in these suspensions!

Next, let’s filter it so that we specifically see rows in which it is likely to be a male character who is being described. Look again in the menu to the right of the screen. Although it might seem more sensible, don’t use the ‘Filter rows’ option at the moment. Look below it to the ‘KWICGrouper’ options.

First, use the slider marked ‘Search in span’ so that you get results from ‘L5’ to ‘R1’. Then, in the ‘Search for types’ box, start typing ‘his’, and when it appears in the drop-down, select it (the way the software works, you have to select an option it has already indexed, so you can’t just type ‘his’ and press enter)

Look both at the actions which are described as being performed by ‘his’ hand, and also at the adjectives which are used to describe the hands themselves.

Now remove the ‘his’ type by clicking on the cross next to it in the ‘Search for types’ box. Replace it with ‘her’, and look again at the actions which female characters tend to perform with their hands and what adjectives are used for them.

Next, compare some of the ways in which characters are described as speaking in short suspensions, (i.e. interruptions to speech which consist of fewer than five words).

Return the values of the options to their original states, although keep ‘DNov’ in the ‘Search the corpora’ box.

Under ‘Only in subsets’ select ‘Short Suspensions’; in ‘Search for terms’ enter, first ‘said’ and have a skim through the results. The results here are, generally, not very interesting!

Try replacing ‘said’ with other options that could be used for expressing speech. Try, for example, ‘shouted’, ‘cried’, ‘sobbed’, and ‘exclaimed’.

Do you get any feel for the way Dickens might be characterising different people by the way they speak? Is there any way in which this might be related to the situations in which they find themselves?

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

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

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

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

*It’s worth playing around with these options a little more. Stylistics rarely yields results on the basis of a couple of searches, and patterns often emerge through forming an intuition from close reading of a text, and then testing it using these kinds of methods, then refining your initial hypothesis on the basis of these results.*

Next, scroll down the right-hand menu options and select ‘Clusters’. This tool will repeated sequences of words (it’s important to note that these are ‘string’ searches rather than lemma searches, which means that ‘waving animatedly’ and ‘waved animatedly’ will be treated as separate items by CLiC).

Select the ‘DNov corpus again; select ‘Long Suspensions’ for the subsets.

The ‘n-gram’ option allows you to select how many words you want to be in your repeated sequences. If you leave it as ‘1’, this is effectively a word frequency count. Set it to ‘3-gram’ and see what the results are. This begins to give you an indication of the ways in which Dickens’s characters are said to behave in these suspensions.

Try different ‘n-gram’ lengths. Be aware that the longer the n-grams and the larger the corpus the tool is searching, the longer the searches will take!

Next, try looking at different corpora. Replace ‘DNov’ with ‘19C’ in the ‘Search the corpora’ box, and select ‘4-gram’ for n-gram length. What are the two most often repeated phrases in long suspensions in the 19th Century literature gathered in this corpus?

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

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

Can you see any patterns in the type of material that appears to be given in 19th C literature long suspensions? Note down the number of instances of some of these (given in the ‘Frequency’ column of the results).

………………………………………………

………………………………………………

………………………………………………

………………………………………………

………………………………………………

Remove the ‘19C’ corpus from the ‘Search the corpora’ box again and replace it with ‘DNov’. How many of the ‘19C’ 4-grams appear in this listing? How many instances turn out to have been from Dickens’s novels? To what does this mean Dickens could be either typical of writing of the time, and to what extent might he be responsible for these phrase-fragments being as prevalent as they are?

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

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

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

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