Methods for Text Mining and Analysis of Text Corpora

Using Word-level Annotation

We have previously looked at text-level metadata, which takes the form of information about the whole text, such as its author, or the date at which it was published, or the place in which it was published. There is a second kind of metadata which may be added to texts at the ‘word-level’, that is to say, information about individual words. We will mostly be looking at automated methods of adding such information, but it is also possible to add user-defined tags for specialised research.

Programmes exist which help users to create and examine this kind of tagging, such as Goldvarb, which is frequently used by linguists. It is also possible to process these user-tagged texts using some programming skills in a language such as Python.

***Parts of Speech***

*It is often very useful when analysing a text to be able to specify parts of speech which you are looking for. These can be used for grammatical purposes, but also for more semantic purposes – that is to say, in further interpreting the views of an author on different subjects, similarly to the way in which we looked at the collocates which are found with certain words. Part of speech tagging is also very useful in analysing collocates.*

Go to the online CLAWS tagger: <http://ucrel.lancs.ac.uk/claws/>

On this page, scroll down to the paragraph with the heading ‘Tagsets’. Find the hyperlink which takes you to the ‘C7 tagset’, and open this in a separate tab so that you can refer to it.

Next, go to the top of the page and follow the link to the ‘Free CLAWS WWW tagger’.

On this page, select the button next to ‘C7’ in the ‘Selected tagset’ options, and ‘Vertical’ in the ‘Select output style’ options. Then enter the following text into the box:

With formidable quickness it moved into the middle of the room, and, as it groped and waved, one corner of its draperies swept across Parkins’s face.

Have a look at the codes which are returned, some selected ones are given below. Look these codes up in the ‘C7 Tagset’ document and note the designation you find for them:

formidable JJ ……………………………………………………………..

it PPH1 ……………………………………………………………..

the AT ……………………………………………………………..

room NN1 ……………………………………………………………..

waved VVD ……………………………………………………………..

its APPGE ……………………………………………………………..

Note that for some of the words more than one tag is given, followed by a number, such as ‘Parkinss [NP1/70] NN1/28 VV0/2’ Can you either figure out or find out what is going on in these situations?

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

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

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

*This type of tagging is conducted on most texts, although they may not always use the same set of ‘tags’ as CLAWS does, depending on the programme used. For example, try searching for the NUPOS tagset and compare it to the C7 tagset. Text analysis tools such as the Historical Thesaurus Semantic Tagger (HTST), which we will use tomorrow, will create versions of text which are tagged in this way, and the HTST uses the C7 tagset. In order to process these, you can load them into an Excel spreadsheet, or you can automatically process the files using scripts that you have written or that have been written for the data format you are using.*

*The easier option for many of these is to find a corpus which has already been tagged with Parts of Speech and provided with an interface which allows you to investigate these. Once again, we will look at one of the BYU corpora to demonstrate this in action*.

***Using Part of Speech Tagging***

*Tags can be used in corpus searches in some interfaces and are especially powerful if they are combined with* regular expressions *or* regexes.

Go to corpora.byu.edu/bnc

Remember you will need to log in to access the corpus for more than a few searches.

Try searching for the word *run*. Are the results which are returned all the same part of speech?

**Spoiler alert** – they’re not. Some will be verbs (e.g. ‘Sit still, Pamela. Don’t run away’), and some will be nouns (e.g. ‘The sun came up about as often as it went down, in the long run’).

Click back on the ‘SEARCH’ option at the top of the page.

Now type *run* into the search box, but click the grey ‘[POS]’ button next to the box. From the drop-down list which appears, select ‘verb.ALL’. Notice that this auto-changes your search term to ‘run\_v\*’. Click ‘Find matching strings’.

Look at the results now, then try to change the search so that instead you get only noun instances of *run*.

Have a good look through the results. Has the filtering been entirely successful? Have you spotted any instances in which verb forms have been misidentified as nouns or vice versa?

Does one seem to be worse than the other?

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

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

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

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

Return to the search option. As well as searching for words tagged as belonging to particular parts of speech, it is possible to search for combinations of words with parts of speech. The BNC tagging uses the C7 tagset which you have looked at above.

Type ‘JJ man’ into the search box. Look at the results which are returned, and click on an option to see its concordance lines. If you want to go back and look at another set of concordance lines, click on the ‘FREQUENCY’ option in the top menu bar. The results are given listed by the frequency of the times they appear. Write down the top five results.

1. ……………………………………..
2. ……………………………………..
3. ……………………………………..
4. ……………………………………..
5. ……………………………………..

 Once you’ve done this, return to the ‘SEARCH’ option again and type ‘JJ woman’ into the search box. Once again, write down the top five results and compare them.

1. ……………………………………..
2. ……………………………………..
3. ……………………………………..
4. ……………………………………..
5. ……………………………………..

Typing a C7 tag in capitals will automatically search for any word with that tag. You may want to try this with other tags which you have found in the exercise earlier in this session. For example, choose a noun and search for that noun preceded by a word with the tag APPGE – e.g. ‘APPGE car’

It is also possible to string these tags together – try searching for a possessive pronoun (APPGE), followed by an adjective (JJ), followed by a noun of your choice – e.g. ‘APPGE JJ car’

***Lemmatisation***

*Some words appear in multiple forms, such as different forms of a verb –* run, runs, running, ran – *or of a noun* – cat, cats, cat’s. *A lemmatiser reduces these to the ‘dictionary’ form, in other words the form that you would expect to find in a dictionary, i.e.* run, cat. *These forms are also known as the word’s ‘lemma’.*

In the BNC corpus, search for ‘run quickly’ and look at the results. Then return to the ‘SEARCH’ option, and type ‘RUN quickly’ into the search box instead. What is the difference in the results that are returned to you?

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

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

Next, return to the ‘SEARCH’ page and type ‘wild cat’ into the search box. Once you’ve done that, return to the ‘SEARCH’ page and type ‘wild CAT’ into the box. What difference do you find?

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

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

Try stringing lemmatization and part of speech together.

* Search for ‘RUN RR’ to get a combination of any form of *run* followed by any adverb.
* Search for ‘JJ CAT’ to get a combination of any adjective followed by any form of *cat*.