BLAXILL AND BEELEN TEXT MINING FOR HISTORIANS COURSE: CLASS 4 EXERCISE

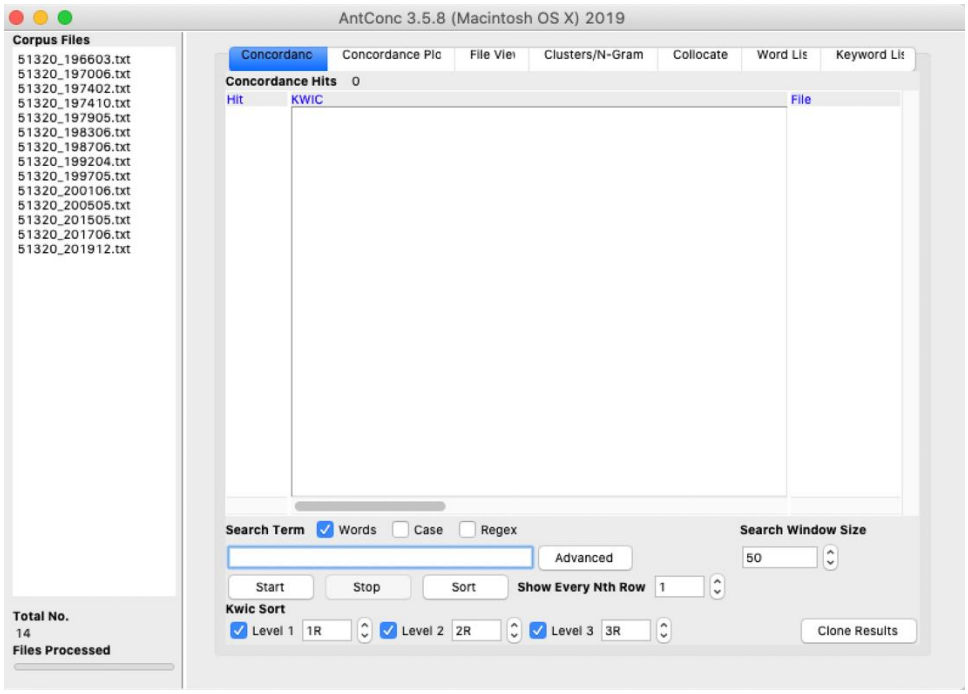
This exercise focusses on the language of party, particularly on how political parties in the UK claim to represent the "people": who are "the people", do Labour and Conservatives claim to speak on behalf of different constituencies? In this exercise we go through the different functions of AntConc and show how each of these contribute to this research question.

Loading the data

We start with importing the Labour Manifestos between 1964 and 2019

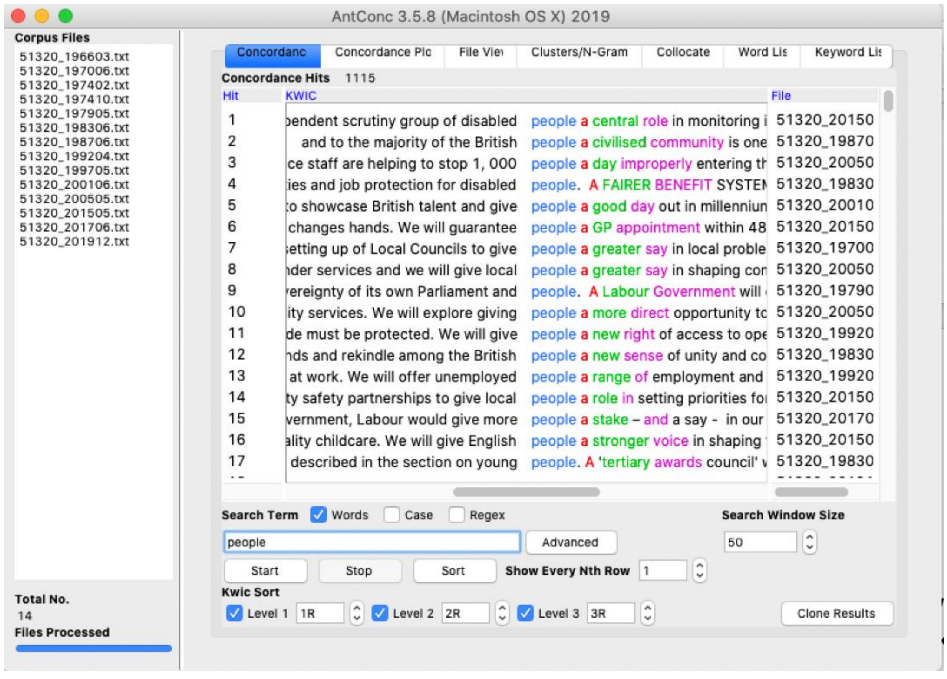
- Download the data from Github: https://github.com/kasparvonbeelen/ghi_python/blob/main/data/manifestos/antconc.zip
- Unzip the antconc.zip folder
- This results in two folders "by_party" and "by_period"
- Open AntConc
- Go File > Open Dir…
- A window pops up that allows you to navigate to the folders you just unzipped
- Select the "by_party" folder, then "labour"
- Click "Choose"

The result should look like this screenshot.



Concordance

After importing the Labour manifestos, click on the “concordance” tab. We proceed by searching for the string “people”, centring the whole collection on this token.

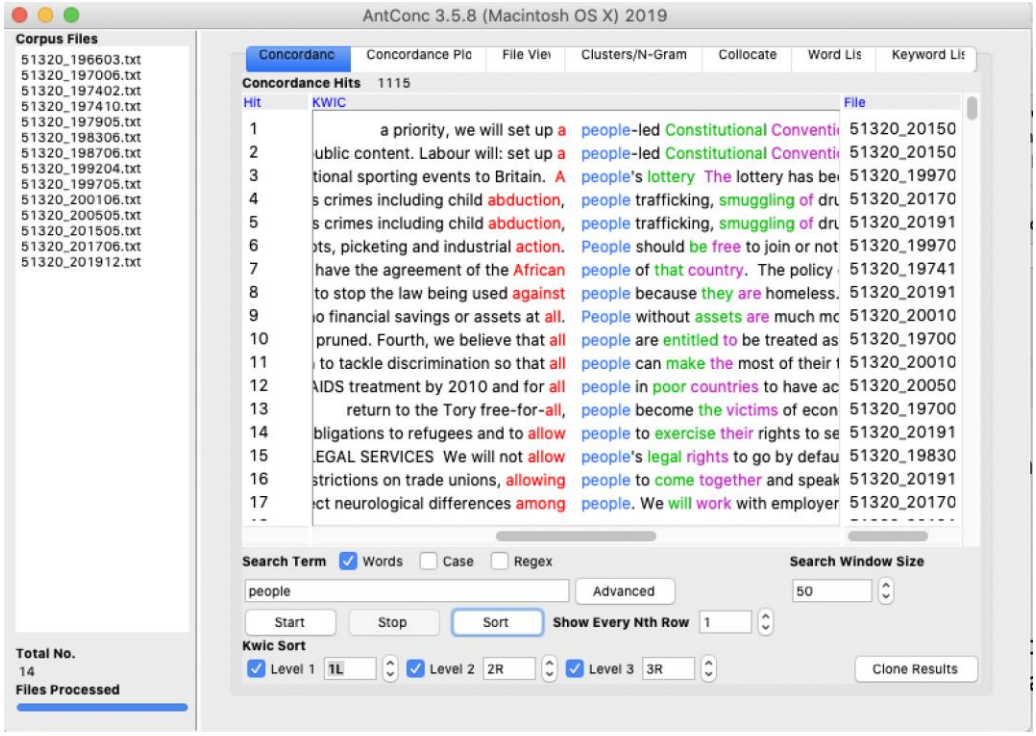


This query shows all occurrences of the token “people” in the centre of the screen. You can increase the context by entering a higher value for “Search Window Size” at the bottom right. You’ll notice that words in the centre have different colours.

- Blue: the query term in the middle

- Red, Green and Magenta are words use for alphabetically sorting the results. At the top of the concordance are all occurrences of the bigram “people a” etc.

Since our query is a noun, sorting results alphabetically on word at the right of our query term maybe doesn’t make a much such, as the adjective qualifying the noun mostly appears on the left. To sort results by the nearest left, look at the “Kwic Sort” panel at the bottom right and change “Level 1” to 1L, then press “Sort”. The result of these setting are shown below, you’ll notice that maybe this a more convenient way to browse the hits for “people”.

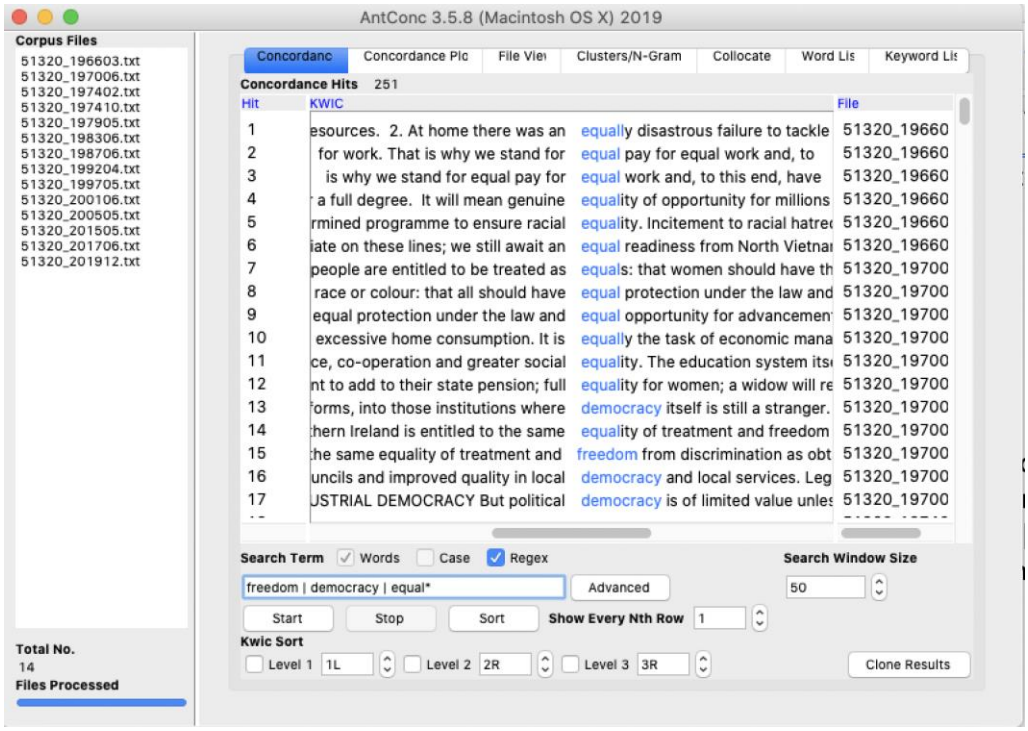


You can experiment more with other sorting methods, for example disable “Level 3” or change “Level 2” to 2L.

To sort the results chronologically, disable all Kwic Sort options. If you look at the file names (the column at the right) you’ll notice they have a party Id (before the “_”) and a date (after the “_”). Since the part Id is the same for all files, the results are then sorted by the string after “_” which is in the format yyyymmdd.

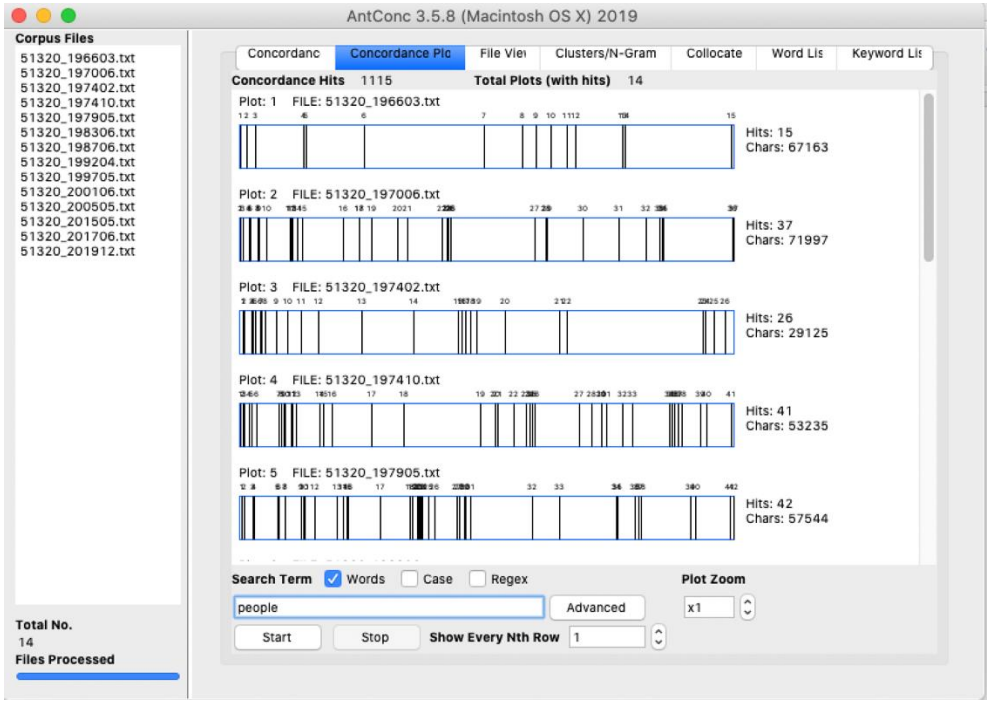
Lastly, you can search for more than one word:

- Tick the “Regex” box next to “Search Term”
- To search for the words “freedom” and “democracy” enter “freedom | democracy” (| is the OR separator) as a Search term (without the quotation marks)
- You can also add a wild card to this search: freedom | democracy | equal*: this finds the tokens “freedom”, “democracy” and all tokens starting with the substring “equal”.



Concordance plot

Whereas “concordance” allows you to zoom in on particular contexts of a word, the “concordance” plots show the query results in a more abstract way: focussing on its location in the corpus. If we search again for the token “people” the resulting concordance plot is shown in the screenshot below.

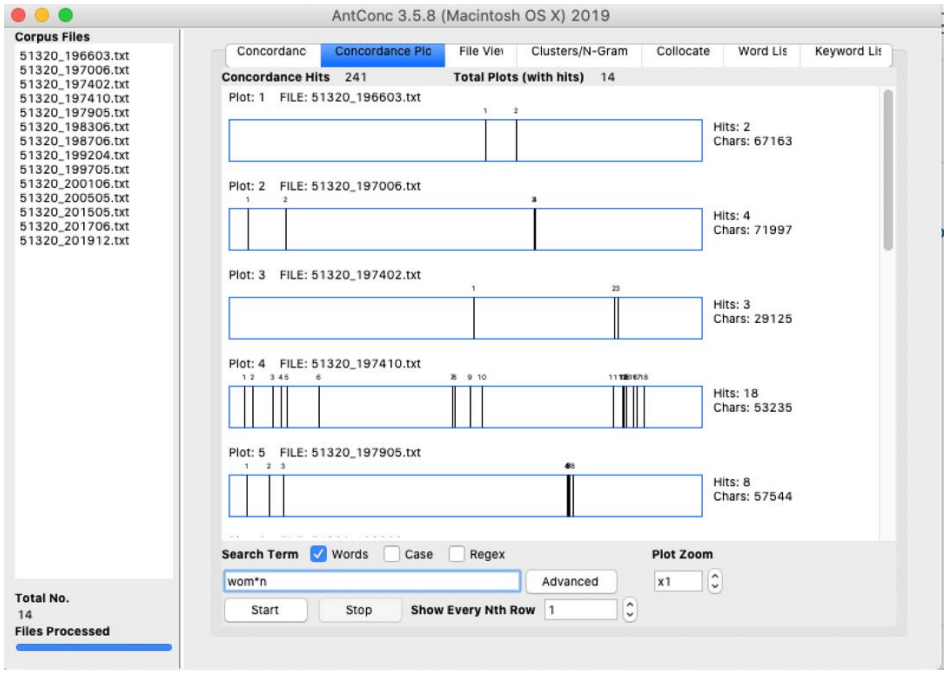


In this case, the concordance plot organizes the query results along two dimensions:

- Horizontal: the plot shows the location of “people” in each manifesto, this visualising if this tokens appears at the beginning or more at the end of each document.
- Vertical: because of the way the documents are named (and the fact the we only have one party in our set of documents) the results appear in chronological order

Such an abstract view on the corpus, could help with comparing the importance of different topics by scrutinizing the both the amount of attention as well the location (does the topic appear on the first pages, or only later in each manifesto.)

For example, compare the query results for “people” with “wom*n” (matches both “woman” and “women”. Result for the latter query is shown below.

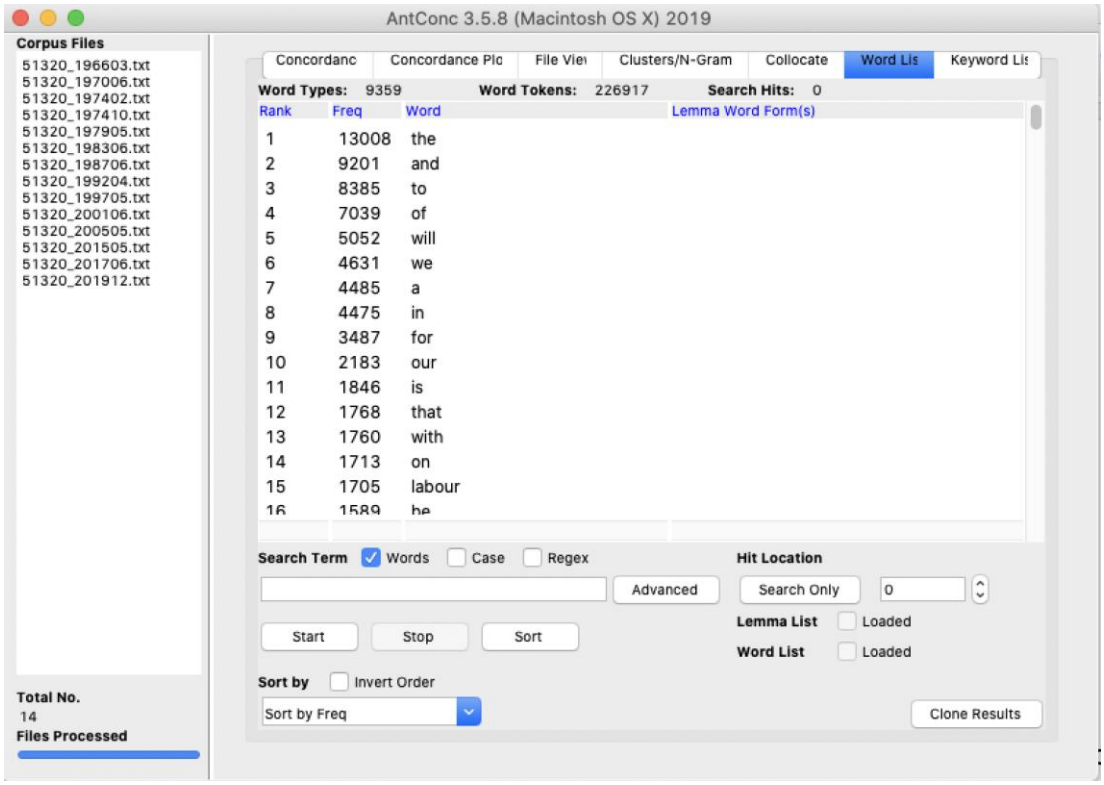


Word list

To get an overall sense of the word frequencies

- Go to the “Word List” tab
- Click “Start”

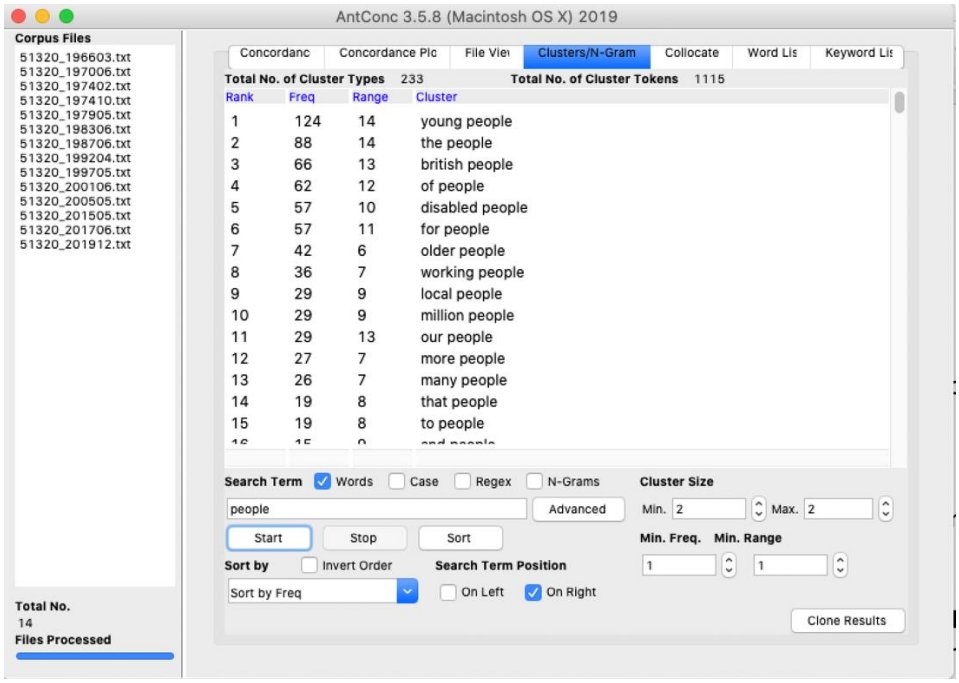
This shows word types in our corpus, ranked by their overall frequency.



Cluster/ngram

The “cluster/ngram” tab computes the frequency of ngrams, a sequence of contiguous words.

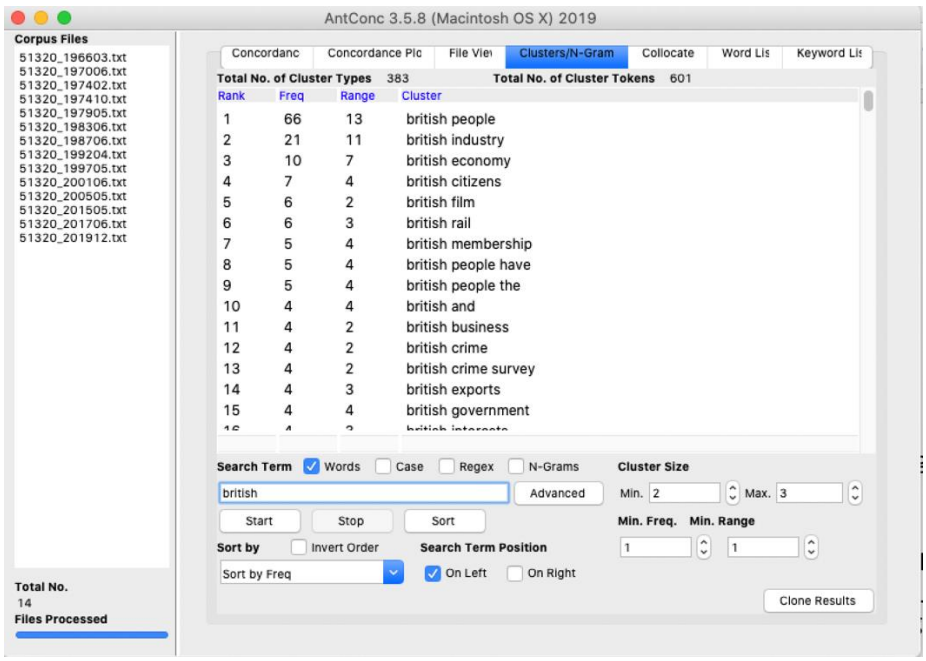
- Enter “people” as “Search Term”
- Inspect the results, you’ll notice that these ngrams are not very interesting, especially the most frequent ones (e.g. “people with” and “people and” don’t tell us much about the context of use.
- Look at the bottom of the AntConc screen, you’ll notice “Search Term Position” is set to “On Left”, change this to “On Right”. This will tell AntConc that our query term must appear at the end of the bigram. You’ll notice that this is a more fruitful way to explore the use of the word “people”: meaningful bigrams such as “young people” and “older people” appear and give us a better sense whom the Labour part is claiming to represent.
- You can easily travel from one tab to the other, for example clicking on “young people” will show you the concordance of this expression, and allow you to study it in more detail. Clicking on a line in the concordance tab, will bring you to the original document.



You can search for longer ngrams:

- In “Cluster Size” change “Max” to “3” this will show ngrams of length 3, also called trigrams.

It is important to adjust your search strategy to the query term. For example, if we are interested in claims around national identity, and want to focus on the adjective “british” change “Search Term Position” to “To Left”, which results in the following trigrams.



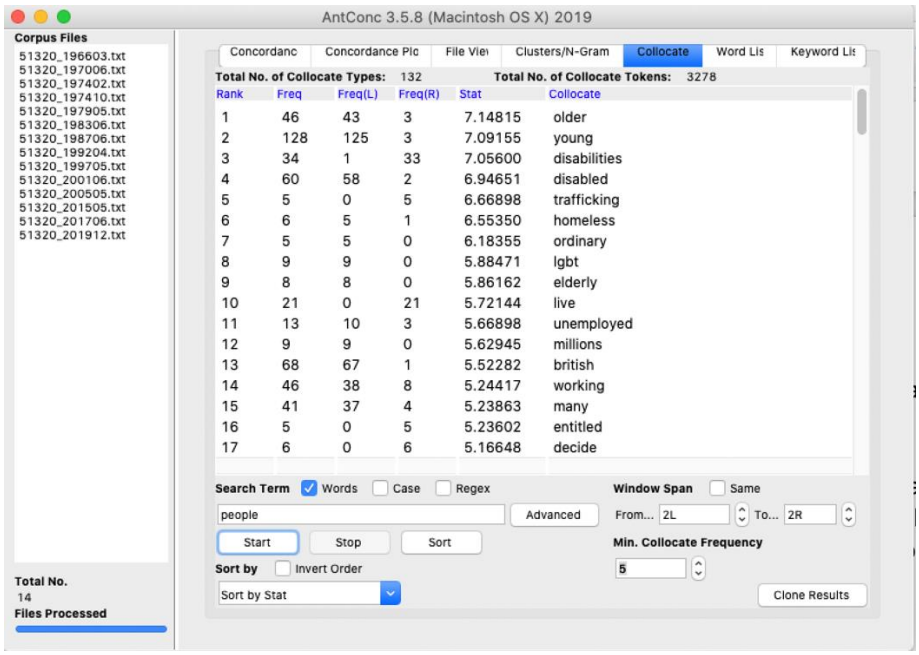
Collocates

The collocates tab allows us to compute which words are “attracted” (according to some measure of association) to our search term. While AntConc makes computing collocations relatively easy, be aware (and make note!) of the specific setting you use, as results may vary depending on the threshold you select.
While there is no right way, there certainly are wrong ways. Most importantly, make sure others will be able to reproduce your results.

Let’s first produce collocation with the standard settings. Enter “people” as Search term and press “Start”. These results look rather strange, why? You’ll notice they all have a very low Frequency (see “Freq.” columns at the left). If two hardly appear in a corpus (i.e. only once), but when they do, they appear close to each other, the collocation score will be high, but this is nonetheless more a spurious association, not a strong pattern.

To produce more meaningful collocation:

- Keep “people” as search term
- Set “Min Collocate Frequency” to 5.
- Make the context a bit smaller: In “Window Span” set “From…” to “2L” and “To…” to “2R” (we only take words into account that appear close to the target word.)
- Press “Start”



However, there are still many more options, you could

- increase the window size, to obtain collocation spanning, for example 10 words: set “From…” to “10L” and “To…” to “10R”

- change the association metric for computing collocations: go to “Settings” and select “Tool Preferences”, which opens a new window. At the left-hand side-the “category” part, select “collocates”. This shows all the other options for computing collocations. Under “Statistics Options” change the “collocation measure” to “T-score”. Click apply, this should

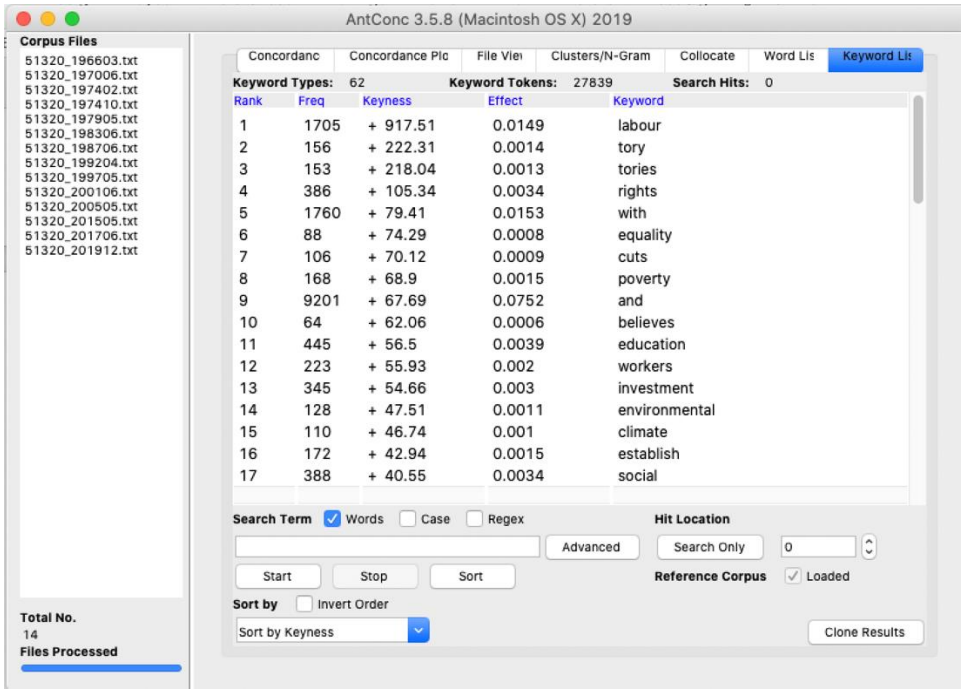
bring you back to the “collocations” tab. Press “Start”, now results look different again, and more confusing, as the statistical measure ranks function words like “the” and “and” very high. Clearly this isn’t the best approach, but it shows how results can very wildly depending on setting and threshold. It is up to the research to defend and document these decisions.

Keyword List

In the last part of this exercise focusses on obtaining words that are characteristic for a certain set of documents (compared to another sample of documents). We will compare which words are indicative of Labour as opposed to Conservative language. First we have load the Conservative manifesto as a “Reference Corpus”:

- In “Settings” got to tool “Tool Preferences”, and click on “Keyword List”
- In the middle, you should see a button “Add Directory”, click on this and navigate to the place where you unzipped your documents and select the “conservative folder” (press “Choose”).
- Once you are back in the “Tool Preferences” window, click “Load” (once these documents are loaded the box next to the “Load” button is ticked.)
- Click “Apply” and go back to the “Keyword List” tab in the main menu.

To compute the words indicative of Labour manifesto just press “Start”. Interestingly, as the screenshot shows, “labour” and “tory” appear at the top.



To compute keywords for the Conservative manifesto’s

- Go back to “Settings” > “Tool Preferences” > “Keyword List”
- Press “Swap with Target Files” and then “Load”
- Press “Apply”
- Go to “Word List” and press “Start”

- Then go back to the “Keyword List” tab, press “Start”