

Introduction to working with Wordsmith Tools at the University of Birmingham

This worksheet provides a basic introduction to using Wordsmith Tools (written by Mike Scott) on the university network. It begins with information about accessing the programme (you must be on the university network to do this, and you must have a university username and password). It then explains the main functions in the tools included in the Wordsmith Tools suite of programmes. It is not a comprehensive guide – if you find that there are features of WST that are not explained here, make use of the Wordsmith Help files.

1 Getting connected

The programme is installed on the university network. When you connect to the programme, a set of documents will be installed in a folder in your U drive (this is the user drive on the network which contains ‘My documents’ and your other personal files). To connect to the programme:

- Open **My computer**
- From **Tools**, choose **Map Network Drive**
- Select **W** as the drive, then type [\\caldfs\cal\Other\Deploy\Software\WSmith5](#) in the path line.
- Then click on **Finish**

The Wordsmith folder should be open on your desktop. Select the icon for **Wordsmith.exe**, right-click on it and choose ‘Send to’ + ‘Desktop (create shortcut)’. This will place a link to Wordsmith on your desktop.

2 What is Wordsmith Tools?

WST is a suite of corpus handling and analysis tools. The main tools are:

- Concord
- Wordlist
- Keywords

In addition to these three tools there are 11 utilities. These are explained in the Wordsmith Tools Help file. We will not have time to look at them today.

Wordsmith Tools can be used to investigate collections of textual data that you have assembled yourself. The data can be simple text or they can be simple text files with annotation added in within tags (for example, <s n=”1”>). The files that you use should be text files (eg, files that end with .txt and also files ending .htm or .xml, etc) and should not be in MS Word format.

In this worksheet, you will learn the basic operations of the three main tools. First, you need to download some corpus data. You will download various collections now, for use in the worksheet.

Open a browser and go to <http://paulslals.org.uk/ccr/wst/>. There are four links on the page. On each link, right-click and choose 'Save Target As' or equivalent) and save to your U drive.

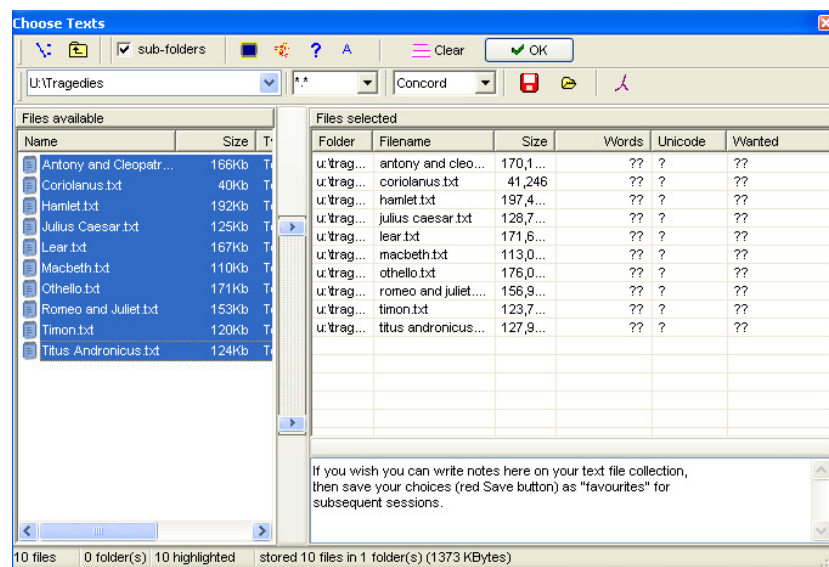
When the downloads are completed, close the browser and open your U drive folder. Two files are zipped (trags.zip and anyqs.zip). Extract all the files from these two zipped files into folders called 'Tragedies' and 'Any Questions'. The first folder contains all of Shakespeare's tragedies, and the second contains transcripts from the BBC political debate programme 'Any Questions'. The other two files contain speeches by Tony Blair from 1997 and from 2005.

3 Concord

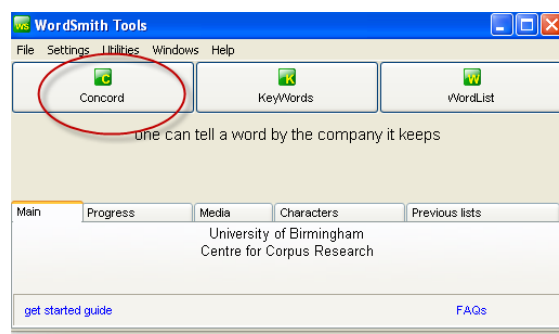
Before you can use any of the tools, you need to 'load' your corpus. This means that you direct the programme to a set of files that you are going to analyse. In this section, you will learn how to load the corpus files, and then how to use the Concord tool.

3.1 Open WordSmith Tools

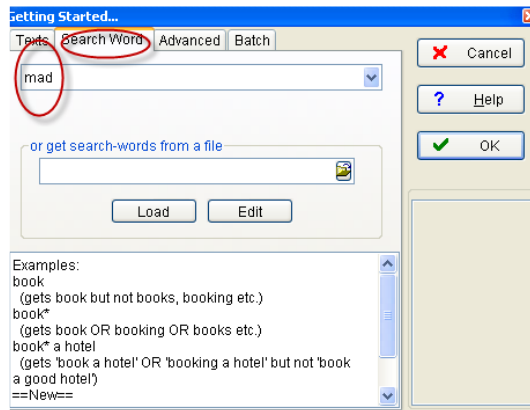
Click on **File** then **Choose texts**. You need to navigate to the U drive (see screenshot below) and open the 'Tragedies' folder, by double-clicking on it. Click on the first file name (Antony and Cleopatra) then press the **Ctrl + A** keys (to Select All the files) and then click on the bar in the middle of the column dividing 'Files available' from 'Files selected'. You should see all the file names appear in the 'Files selected' area. Now click on **OK**.



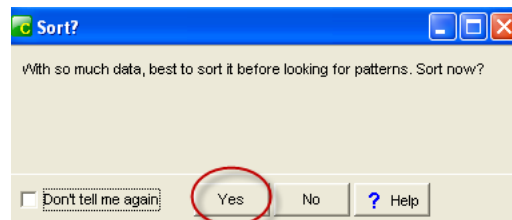
Click on the Concord button to open the Concord Tool.



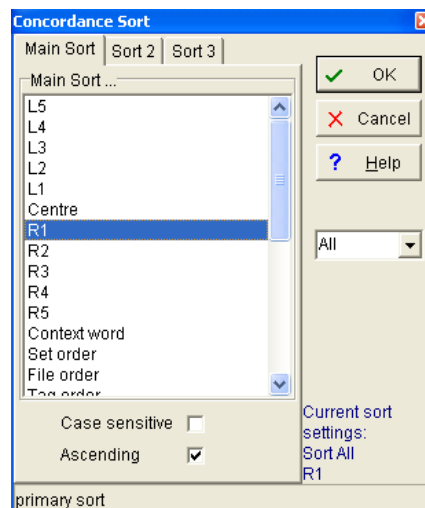
Choose **File** then **New**. In the Getting Started box, under 'Search Word' specify the search word as 'mad':



You will be asked if you want to sort the data. Choose 'Yes':



This time, click on **R1** and then **OK**. This means that you are choosing to sort the data by the first word to the right of the search word. There are many other options – you can choose to search the third word to the left, for example, which is L3 – and you can also specify the first criterion for sorting (L1, for example) followed by the second (for example, L2) which has to be specified under the 'Sort 2' tab.



Click OK and then you will see the concordance window:

If you double click on any line, you will go to the relevant part of the source text – you can then read the wider co-text of the line. To return to the concordance view, click on the ‘concordance tab’ in the bottom left corner.

Madness is a common theme in Shakespeare’s tragedies. Let us now search for variants on the word ‘mad’. Choose **File** then **New**. When asked if you want to start a new window, say ‘Yes’ and then, in the new window, once again choose **File** then **New**. In the ‘Getting started’ box, type:

Mad/ madness/madman

Return to the ‘Concordance’ view and sort the data by the centre word this time. Drag the ‘Set’ column heading to the right to make the column wider.

How many times does each word occur in Shakespeare’s tragedies? The easiest way to find this out is to use the ‘Collocates’ tab, and click (twice) on the ‘Centre’ column heading.

Mad
 Madness
 Madman

Resort the data by ‘File’. In which play is there the most reference to madness?

.....

How often does each word occur in each play and where in the play is each word used? To answer this, look at the ‘Plot’ tab. This shows the hits for each word in each file and also gives a visual guide to where the word occurs in the file.

3.2 Activity

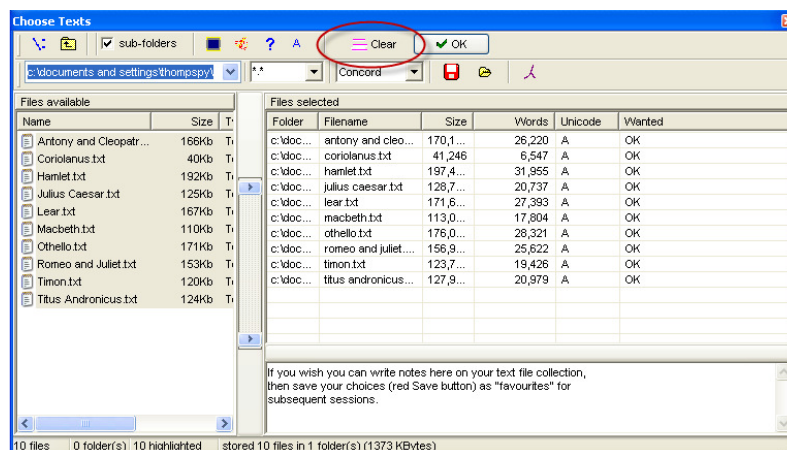
Another common theme in the tragedies is that of fortune, or fate. Investigate the uses of ‘fortune’, ‘fate’, ‘fates’ and ‘destiny’ in the tragedies. Work with another student and write a summary of your observations in the space below:

3.3 Coding and deleting lines

In the previous investigation, you will have seen that the ‘Set’ column showed you which of the search words appeared in the particular line. You can use the ‘Set’ column for another purpose – to code your data.

You are going to look at the language of Blair’s speeches, in 1997 and in 2005, and the purpose of the investigation is to find out how the language of his speeches changed.

You need to load a different corpus now. Go to the **File** menu and choose **New**. In the ‘Getting started’ window, click on the ‘Texts’ tab, and then ‘Change selection’. IN the ‘Choose texts’ window, click on the **Clear** button (next to the **OK** button).



Navigate back to the U drive and select the **blair97.txt** file. Click **OK**.

Type ‘we’ into the search term box. Then click **Start**. Sort the lines by R1.

Which modal auxiliaries are used and how many times for each?

Now change your choice of text. Close the **blair97.txt** file and open **blair05.txt**. Repeat the query.

Which modal auxiliaries are used and how many times for each?

What does the difference in the use of modal auxiliaries tell you about how Blair's discourse has changed?

Coding lines

Sometimes when Blair says 'we' he is referring to the Labour Party, and sometimes to the Cabinet, and sometimes to the country at large (we Britons ...). Look at the first 20 lines and decide whether which of these meanings of 'we' is shown. Use the code 'A' for Labour, 'B' for the Cabinet and 'C' for the country. If the 'we' in the line is Labour, type the letter into the Set column in that row. You can code each line by typing a letter into the column.

You can resort the data by the codes – click on the heading for the 'Set' column.

Investigating adverbs

The texts that you are looking at are not tagged for Part of Speech. In the next activity, however, we are going to see how it is possible to investigate the behaviour of a particular part of speech without tagging. In English, adverbs usually end with the letters L and Y.

In the Search Word box, type ***ly**, And then click on **'Start'**

[the asterisk is a wild card – it will capture any letters before the final 'ly'. This is a very useful expression to use]

Sort the lines by the Centre word. There are some hits that are not adverbs – you can get rid of these lines in the following way:

1. Select the line, and then press the 'Delete' key – the line will change to a grey colour and will be struck through
2. When you have decided on all the lines that you want to remove, go to the 'Edit' menu and choose **Zap**. The lines will disappear.

After deleting lines, you will need to compute things like 'collocates', 'plot' and 'clusters' again. In the 'Concordance' view, go to the **Compute** menu and choose the feature that you are interested in.

3.4 Activity 2

Which are the most frequently used adverbs, in the 1997 data and in the 2005 data? What do you notice about these four adverbs? What are they used for (what is the rhetorical purpose)? Write your notes in the box below:

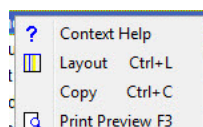
Do you notice any differences between the uses of adverbs in Blair's 1997 speeches and his 2005 speeches? Write your ideas down in the box below:

Comparison

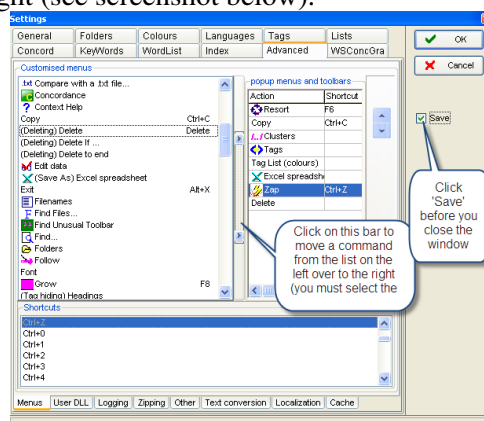
3.5 Saving data

You can save your concordance data in four formats: plain text, as XML, as an Excel spreadsheet or as a .cnc file (this can only be opened in WST). Note that you also open XML files in Excel, as well as opening in an XML editor or in a text editing programme.

Tip: If you use WST a lot, you will find that there are certain commands that you use frequently. It is handy to have these commands available in the right-click menu (the 'Context' menu), and WST allows you to customise the context menu. To do this, go to **Settings** then **Customise**. You can remove commands from the menu (shown below) by selecting them and then pressing the **Delete** key.



You can then add new commands to the menu by selecting from the left side list and transferring them to the right (see screenshot below).



4 Wordlist

You can create a wordlist for a given group of files. There are two ways to do this in WST – one is to create the list on the fly, as it were, and the other is to index the corpus and then derive the wordlist from the index. You will first look at how to make a wordlist on the fly, and then secondly look at creating an index for your corpus. What's the difference? If you have an indexed corpus, you will be able to do some more sophisticated investigations of your data, particularly the n-grams in your corpus. If, however, you are not interested in these types of analysis, you do not need to index your data and you can save yourself time.

4.1 Making your wordlist

To create a wordlist, as usual you have to load your corpus first. Open WST and choose the Wordlist tool. Go to File and choose New. When prompted to choose your texts, locate the Shakespeare tragedies and select all of the files. Then click on OK.

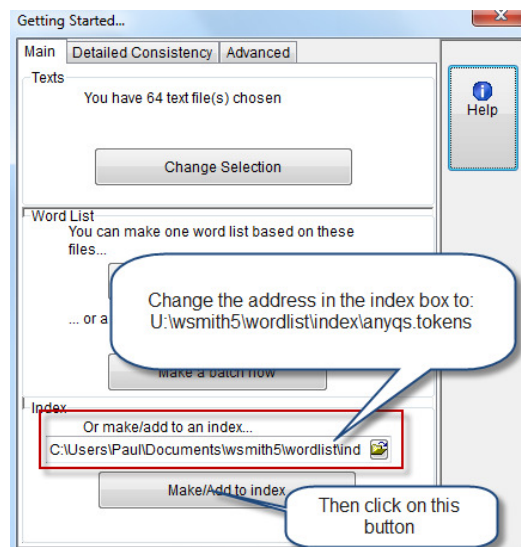
You will then be asked whether you want to make a single wordlist, a batch or whether you want to build an index. In this instance, choose to make a single file.

When the operation has completed, you will have a window in which you can see the wordlist sorted by frequency. At the bottom of the screen are tabs for accessing the same data sorted alphabetically and also a summary table of the statistics for the wordlists.

Make one wordlist for all the tragedies and call this 'tragedies' and then make a second wordlist, using ONLY the Othello file and call that 'Othello'. You will use these files in 5 below.

Now let's try using the indexer. Follow these instructions carefully:

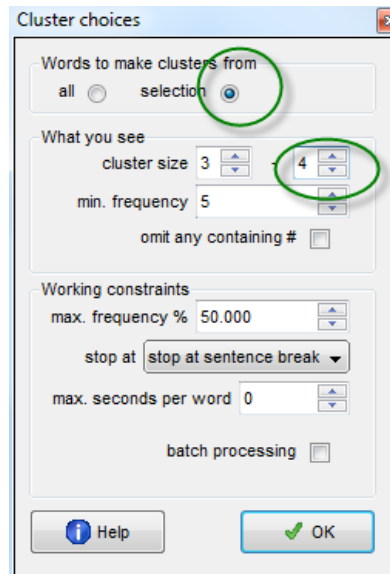
Create a new wordlist, and choose as your texts the files in the 'anyqs' folder.



This process will take a long time. When it has finished, you will see the wordlist windows as usual. One thing that you can do with an indexed wordlist is to investigate clusters.

You can compute the clusters for all the words in the list but this will take a very long time with a long word list like this. Instead, you are going to select one word: **actually**. To select it, click on the word in the wordlist and then hit the F5 key to 'mark' the line.

Now go to **Compute** and choose **Clusters**. In the new window, make two changes:



You should now see a listing of the 3 and 4 word clusters involving 'actually'. Make a note of the clusters, and now repeat the process to see what clusters there are for 'really'. To return to the full list of words, go to **File** and **Revert to index**. Make a note of the clusters below:

actually	really

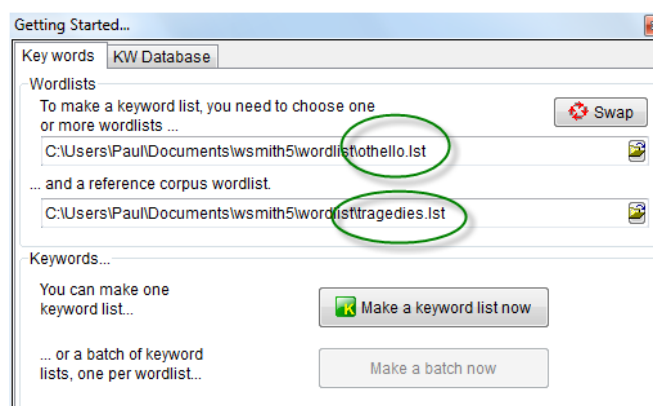
5 KeyWords

The KeyWords tool allows you to identify tokens which occur with greater relative frequency in one set of data than in another. Typically, a keyword analysis is conducted by comparing the word frequencies in one set of data (the data that you are primarily interested in) with another set of data, much larger, which is called the reference corpus. The assumption here is that you are determining how language is used in your corpus compared to language in a more general dataset. Some researchers choose to use a wordlist of the British National Corpus for the reference corpus – in this worksheet we will use a different reference corpus (to save time) but you can download the BNC wordlist from Mike Scott’s website, if you want to make use of it another time:

<http://www.lexically.net/wordsmith/>

Before you can make a keyword analysis, you need to make wordlists for the two corpora that you want to compare. In the previous section you made one wordlist for all the Shakespeare tragedies and another for the play ‘Othello’.

Return to the initial Wordsmith window, and choose ‘KeyWords’. Then **New**.



In the first line, choose the **Othello** wordlist by browsing to the ‘wordlist’ folder in the ‘wsmith5’ folder [click on the folder icon at the end of the line to start browsing]. In the second line, for the reference corpus choose the **tragedies** wordlist. Then click on ‘Make a keyword list now’. The keywords in red at the bottom of the list are ‘negative keywords’ – they are markedly less frequent in Othello than in the other tragedies. The black words are all positive keywords. What do you learn Othello from the keyword list?

N	Key word	Freq.	%	Freq.	RC. %	eyness	P	emmas	Set
1	CASSIO	113	0.43	113	0.05	210.56	0.0000000000		
2	IAGO	60	0.23	60	0.03	111.72	0.0000000000		
3	MOOR	56	0.21	74	0.04	86.04	0.0000000000		
4	DESDEMONA	40	0.15	40	0.02	74.46	0.0000000000		
5	SHE	157	0.60	563	0.27	67.92	0.0000000000		
6	I	832	3.19	4,948	2.37	60.70	0.0000000000		
7	HER	213	0.82	918	0.44	58.60	0.0000000000		
8	HANDKERCHIEF	28	0.11	28	0.01	52.11	0.0000000000		
9	LIEUTENANT	29	0.11	31	0.01	51.74	0.0000000000		
10	RODERIGO	27	0.10	27	0.01	50.25	0.0000000000		
11	OTHELLO	24	0.09	24	0.01	44.67	0.0000000000		
12	CYPRUS	23	0.09	24	0.01	41.67	0.0000000000		
13	DO	221	0.85	1,094	0.52	38.56	0.0000000000		
14	HEAVEN	62	0.24	197	0.09	33.95	0.0000000027		
15	WILLOW	18	0.07	19		32.37	0.0000000098		
16	VENICE	17	0.07	17		31.63	0.0000000157		
17	HONEST	42	0.16	117	0.06	28.65	0.0000000838		
18	WE	52	0.20	865	0.41	-32.73	0.0000000077		
19	OUR	51	0.20	896	0.43	-37.91	0.0000000000		