CORPUS LINGUISTICS
Wmatrix Practical: Introduction to Wmatrix

1. Site orientation (folders, different windows and what you can do with Wmatrix)

1.1 Logging in
To access the Wmatrix environment you need to type the following URL into your web-browser: http://ucrel.lancs.ac.uk/wmatrix3.html

You should be presented with a login box, into which you type the username and password that you have been given for this session. Then click on OK.

***NOTE THAT FOR CLASS SESSIONS THE WMATRIX USERNAMES AND PASSWORDS ARE SHARED BY OTHER STUDENTS, SO PLEASE DO NOT DELETE ANY FOLDERS. ***

1.2 Folders etc.
When you log in to Wmatrix, you will see the following ‘Welcome’ page:

Click on ‘My folders’ – and you will see a page similar to the one below:

To see the folders loaded into your work area, click on the ‘My folders’ link.
You will see a number of folders (denoted by the 📁 icon) in your Wmatrix user area. When you load a file into Wmatrix a folder is created automatically to hold that file. It is best to load each data file into its own folder. Each folder icon forms a link to a page showing the contents of that folder.

Click on the folder called (’EMAIL’), to see the ‘file view’ screen similar to the following:

The screen might look slightly confusing at first, as there are many different options which can be selected.

The screen is divided into a number of parts. The top of the screen, between two horizontal orange lines, contains a menu containing links to functions which are mainly administrative, e.g. loading files or deleting folders. These functions appear in this menu format at the top of all the Wmatrix pages you will visit. Below the admin links is a table-like section, which displays options for viewing and comparing the data loaded into the folder.

Looking first at the admin menu part of the screen, there are four groups (denoted by bold text): ‘Tagging’, ‘Folders’, ‘Options’ and ‘Help’. Each group contains a number of links (denoted by blue text), which perform different functions.

For example, if you click on ‘my folders’ in the Folders group, this will return you to the screen showing all the folders currently loaded in your folder.

Click on the EMAIL icon again to get back to the ‘folder-view’ screen.

Moving on to the table-like part of the screen, you will notice that the top part of the table is divided into vertical columns and horizontal rows. The three rows (‘Word’, ‘Part of speech’ and ‘Semantic’) relate to the types of lists you will be able to view. So, the ‘word’ row will show you lists and perform functions at the word level, the ‘part of speech’ row performs functions at the grammatical level, and the ‘semantic’ row at the semantic level. The columns give you various options for viewing or comparing output.
Word frequency

Starting with word lists, it is possible to see these lists presented in two different ways: sorted by frequency, or sorted by word (i.e. alphabetically).

☐ Click on the ‘Frequency’ link in the ‘Frequency list’ column of the ‘Word’ row.

You will see a list of all the words contained in the corpus in order of frequency, with the most frequent words at the top of the list.

The column headings (‘word’ and ‘frequency’) are also links.

☐ Click on ‘word’ and the screen will refresh, showing you the words in the text sorted in alphabetical order, with their frequency in the right hand column.

☐ Click on ‘frequency’ to return to the list of words sorted in order of frequency.

The third column (‘Relative frequency’) shows the percentage frequency of the word (the frequency divided by the total number of words in the corpus multiplied by 100).

Concordances

The final column of the table, which has no heading, has links to concordance data for the corresponding word.

☐ Click on the ‘concordance’ link for ‘I’ to display the following screen:
You can widen or shorten the length of the concordance lines displayed in the list by adjusting the ‘Change character width’, then clicking on ‘Go’.

You can also view more of the context for one result at a time, by clicking on ‘More’ or ‘Full’. Notice that when you do this, the extended concordance opens up in a new window or tab in your browser. To get back to your lists, you must go back to your original browser window or tab.

To return to the list of words click on the ‘Back’ button of your internet browser, or

To return to the folder view, click on the ‘EMAIL’ link in the ‘You are here’ portion of the admin portion of the screen.

In the Folder view, if you now click on the ‘word’ link next to the ‘frequency’ link you clicked on earlier, you will see that this also takes you to the word list in alphabetical order.

Now return (again) to the ‘folder view’ screen for this folder.
Moving on to the other types of lists, you will notice that the POS and Semantic areas of the table show two ‘sorted by’ options: tag + frequencies; and word + tag + frequencies. One presents three columns showing tag information in one column and raw and relatives frequencies in the other two, while the other presents four columns, showing words, tags and frequencies. We will look at these screens now, but we are going to skip POS and deal only with the ‘semantic’ part of the table.

(i) Clicking on this link will give you a three-column list containing semantic tag, raw and relative frequencies.

(ii) Clicking on this link will give you a four-column list containing words, semantic tags and raw and relative frequencies.

Click on the semantic ‘frequency’ link that will give you a three column list – (indicated above), and you will see the following screen:

This screen shows the semantic tags listed in order of frequency. So Z5 is the most frequent semantic tag in the corpus. You will notice that two frequencies are given:

- **Click on ‘Semtag’ to see the list ordered alphabetically by semantic tag:**
- **Click on ‘Concordance’ to see all the words with the Z8 tag as they occur in the corpus:**
- **Click on ‘List’ to see a list of all the words with the Z8 tag:**
‘raw’ frequency – the total number of words in that semantic category;
relative frequency – the total number of words in the semantic category divided by the total number of words in the corpus multiplied by 100.

By using the various links on this screen it is possible to see other information concerning semantic groups and the words within those groups.

Click on the ‘EMAIL’ link toward the top of your screen in the ‘You are here’ part of the screen to go back to the folder-view screen.

Now click on the semantic ‘frequency’ link that will give you a four column list and you will see a screen similar to the one below:

<table>
<thead>
<tr>
<th>Word, semantic tag and frequencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorted on frequency.</td>
</tr>
<tr>
<td>You are viewing a frequency profile. Click on a column heading to sort on that column. Click on a Concor dance link to see concordance lines. Please note that concordances are not filtered by tags, so you will contain all occurrences of the word.</td>
</tr>
</tbody>
</table>

Search shortcuts: Show complete list | Go |
Search this list: Enter the word or tag you wish to search for here: Go (you can also search for part of a word or tag: enter ‘?’ or leave blank for complete list) |

This table shows all the words in the text in order of frequency with the associated semantic tag. Again, it is possible to see this list of words in alphabetical order by word and by semantic tag by using the links at the top of the columns. It is also possible to see the words as they appear in the corpus data by using the concordance links associated with each word.

There are links to the lists of POS and semantic domain tagsets on the menu at the top of the screen.

2. Making comparisons

Wmatrix allows you to compare your text/data with other data in terms of:

(i) words,
(ii) parts of speech, and
(iii) semantic groups.
This means that you can compare the word list for a text with the word list of, say, a larger corpus of data. The differences between the relative frequencies of words in the texts are tested for statistical significance using the log-likelihood (LL) calculation. This results in a list of ‘key words’, with the most statistically significantly overused words at the top of the list. Similarly, this process can be done for parts of speech (to give ‘key POS’) and semantic domains (groups) (to give ‘key concepts’).

We will first make a key word comparison, comparing the words in EMAIL with the words in the BNC Written Sampler corpus (1 million words from BNC written corpus), which is a corpus file already loaded into Wmatrix.

2.1 General discussion – what basis/bases of comparison to use?

There is an issue, of course, regarding which texts/data sets to compare particular texts with. Should we compare EMAIL to:

(i) the whole written part of the BNC sampler corpus – which contains a variety of written genres;
(ii) the spoken part of the BNC sampler, or
(iii) a specific genre represented by a sub-section within the corpus?

2.2 Comparing the word frequencies in EMAIL with those of the BNC Written sampler.

Go to the folder view of the ‘EMAIL’ folder

Click on the down arrow of the drop-down-menu box in the ‘word’ row of the table. This should present you with a list of possible files with which to compare the wordlist for the whole corpus.

Select ‘BNC Sampler Written’, and click on ‘Go’.

2.2.1 Lists

When you click on ‘Go’, you will be presented with a screen that contains a list of ‘Key words’. These are words that appear more in EMAIL than in the comparison text.
The list has eight columns:

1) Concordance link – click on this to see a concordance list for each word as it occurs in EMAIL
2) Item - the word item
3) 01 - The raw total for that item in the EMAIL corpus (text 01)
4) %1 - The relative or percentage frequency of that item in EMAIL (% in text 01)
5) 02 - The raw total of the word item in the BNC Sampler (text 02)
6) %2 - The relative frequency of that item in the BNC Sampler (% in text 02)
7) A + sign that denotes that the word item appears relatively more in text 01 than it does in text 02
8) The log-likelihood (LL) score – a calculation of statistical significance.

Wmatrix shows you all the overused words with a LL score right down to zero. It is up to you to decide at what point in the list to stop looking; in other words, you need to decide a cut-off point for your analysis. This might relate to a level of significance (a LL value) (see more about LL here http://ucrel.lancs.ac.uk/llwizard.html).

2.2.2 Word Clouds

Use the side scroll bar to scroll down to the very bottom of the key word list. You will find a ‘word cloud’, which shows the most significant key words in alphabetical order. The size of the word in the cloud relates to its LL score, so the bigger the word the more significant or ‘key’ it is (see below).
Move your cursor over a word, and a text box will appear containing the raw frequency and the LL score for that word.

Click on a word to see the concordance lines for that word.

Word cloud

2.3 Comparison of EMAIL with the BNC spoken sampler.

In this part of the workshop we will be comparing EMAIL with the BNC spoken sampler. Practise going through the steps to make the comparison. How do the results differ from when you compared EMAIL with the BNC written sampler? Does this tell you anything about the EMAIL corpus and/or the language used in emails?

3. Uploading a file (using Tag Wizard)

To upload a file into your folder and tag it, use the ‘Tag Wizard’ facility.

Click on the ‘Tag wizard’ link toward the top of the screen in the admin menu area (shown below) to get to the Tag wizard screen.

Tag wizard link

*** NOTE: Files loaded into Wmatrix must be PLAIN TEXT (.txt) format ***
In the tag wizard screen enter the name of the folder in which you wish to store the uploaded file and its associated wordlists (box 1). This can be any name, so choose something that helps you remember what is in that folder.

Next using the browse button to select your file (which will automatically fill in box 2).

Now click on the ‘3. Upload now’ button.

Your text file will be uploaded into your Wmatrix folder with the name you specified in box (1). When Wmatrix uploads your file, it automatically tags it for parts of speech (POS) and semantic categories (USAS). During the upload process (which might take a number of minutes) you will see the following screen:

When the upload is complete, you will be taken automatically to the folder view of the new folder.

6. Time to Explore

If there is time left, you can explore Wmatrix in any way you like. On the Wmatrix home page [http://ucrel.lancs.ac.uk/wmatrix/](http://ucrel.lancs.ac.uk/wmatrix/) there is useful information about the programme, including how to cite and reference it, plus links to some helpful mini-tutorials [http://ucrel.lancs.ac.uk/wmatrix/tutorial/](http://ucrel.lancs.ac.uk/wmatrix/tutorial/). There is also a useful ‘Help’ section [http://stig.lancs.ac.uk/cgi-bin/wmatrix3/help.pl](http://stig.lancs.ac.uk/cgi-bin/wmatrix3/help.pl).