# Table of Contents

**Foreword** 0

**Part I** Introduction 2

**Part II** Choosing your texts 4

**Part III** Select the right language 8

**Part IV** Concordancing 12

1 overview .......................................................... 12
2 making a concordance ........................................ 13
3 seeing the source text ....................................... 15
4 collocates and mutual information .................... 17
5 concordancing tagged text (1) .......................... 20
6 concordancing tagged text (2) ........................ 21

**Part V** WordList 24

1 overview .......................................................... 24
2 making a word list ........................................... 24
3 concordancing selected words .......................... 27
4 lemmatising ..................................................... 29
5 word list statistics ............................................ 31
6 multi-word units .............................................. 32
    using an index ............................................... 32
    making a multi-word wordlist ......................... 33

**Part VI** KeyWords 35

1 overview ........................................................ 35
2 making a key word list .................................... 35
3 key words plot ............................................... 37
4 concordancing selected key words ................... 38

**Index** 40
Step-by-step guide to WordSmith

Introduction

Section I
1 Introduction

These pages are to help get you started. Screenshots take you through each stage.

This is the main screen of the WordSmith Tools Controller.

It has a saying (which keeps on changing and which you can edit), three buttons for the main Tools (Concord is shown as in use), and a series of tabs. At the moment we see the main one showing that *anthony & cleopatra.txt* has been chosen for Concord.
Choosing your texts
2 Choosing your texts

To choose text files, click the File menu in the main Controller:

When you click Choose Texts, you will see something like this:

At the left is a fairly standard text file explorer, at the right an area for Files selected.
Press the browse button ( ) to find the folder where your texts are. You need plain text (.txt) files.

Click the button with the two small arrows, or drag some text files from left to right. You should see something like this:

![Choose Texts]

drag the files over here to the right
or else just click here

Click the button with the two small arrows, or drag some text files from left to right. You should see something like this:
At the moment WordSmith shows (in the status bar just above) that 6 have been stored. You can see the file sizes but WordSmith doesn't (yet) know how many words there are in each text file. We have chosen 6 texts for Concord (see Concord just above Files selected).

Press the green button or just close the window.
Select the right language

Section III
Select the right language

Most examples in this guide deal with texts in English. If you want to handle texts in Chinese or some other language, you need to choose the language in the main Controller.

If in the drop-down list (showing ) the language you need isn't available, click Edit Languages and choose the language you want:
and drag it to the right or click the button in the middle.

You'll see an option to select that language as your "main language" (with which WordSmith will start up by default), or else just as an available language.
In this screenshot, English has been selected and some suitable choices for English have been made such as apostrophes allowed within a word, hyphens separating forms like self-conscious into two words and showing that Arial 10 is the default font, etc.

Finally, save your settings.
Concordancing

Section IV
4 Concordancing

4.1 overview

A concordance looks something like this:

![Concordance Example](image)

It's a concordance of all the occurrences of *wherefore* in Romeo and Juliet. There are only 5 entries. The famous one comes 6,537 words (27%) into the play.
4.2 making a concordance

When you press the Concord button in the main Controller, a new Concord Tool opens up and will be visible in the Windows Taskbar.

Now in Concord itself, choose File | New.

If no text files have been chosen, you are asked to choose some. Press the Choose Texts Now button.
Once the texts have been chosen, enter a suitable *Search Word*.

Here I have chosen *wherefore* as my search-word. Then press OK.
The concordance lists all the examples of "wherefore" which had a word-separator such as punctuation, space etc. before and after it.

Since we have now done our concordance, WordSmith now knows how many words there are in each text file: romeo and juliet.txt has 24,275 altogether.

4.3 seeing the source text

To see the source text, double-click on the line in question. Here, I clicked on the highlighted line containing wherefore art thou Romeo.
or press F8 and the lines grow fatter:

```
<table>
<thead>
<tr>
<th>Concor...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
</tbody>
</table>
```

or pull the line you're interested in wider or fatter; place your cursor at the bottom the number in the left column, and it changes shape:

```
| 5  | bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy |
```

and pull it down.
You can also pull it wider by putting your cursor at the right edge, just to the left of the word Set.

4.4 collocates and mutual information

Here are the collocates of AGO computed using the written section of the BNC, ordered by frequency.
There are nearly 17,000 instances of AGO, and YEARS is the top collocate, found 9,000 times near AGO. The "Relation" column is blank. What's needed is a way of knowing how closely each of these collocates of AGO is related to it. Are A, THE, WAS etc. really closely linked to AGO?

If we now choose Compute | Relationships in the menu,

and select a suitable word-list to use for the comparison:
then we get the following list when sorted by clicking the **Relation** column:

The top items in the list now reflect much better the tendency of **AGO** to accompany periods of time and numbers. [The top collocates **HENSLEY** and **GROSS** only occur 5 times each with **AGO** but out of small numbers altogether in the whole BNC Written.]
4.5 concordancing tagged text (1)

Probably the first thing to do if your source text is tagged, is to let WordSmith know. To do this, in the main Controller, choose Settings | Adjust Settings

Then Tags.

If you're using the British National Corpus (world edition), choose it within Custom settings as shown above.

So far, we have told the Controller that it is to ignore all tags beginning and ending with angle brackets (< >), to translate a few entity references to symbols like % and ”, and to cut out the header of each text (up to </teiHeader>). That'll do for a start.
4.6 concordancing tagged text (2)

Now, we are going to do a concordance on a part of speech. The BNC uses mark-up like this:

\(<w\ PRP>at \ AT0>the \ AJ0>great \ NN2>houses\)

so each preposition is marked \(<w\ PRP>\) just before the preposition itself. The aim is to see all the prepositions in a selected text from the BNC. With a BNC text file chosen, type \(<w\ PRP>*\) as the search-word (the asterisk is needed because a word follows directly after the part-of speech tag) and press OK.

WordSmith checks whether the angle-brackets are text characters or tag-openers and -closers:

![Confirm window](image)

Here we answer Yes.
You see the prepositions and their tags (but no other tags).
Step-by-step guide to WordSmith

WordList

Section V
5 WordList

5.1 overview

A word list in WordSmith Tools looks something like this:

![WordList screenshot]

It shows how often each word occurs in the text files, what that is as a percent of the running words in the text, and how many text files each word was found in.

5.2 making a word list

To make a word list, first press the WordList button in the main Controller.

![WordList button]

When WordList starts up, choose your texts and then you will see something like this.
Here we're going to make one simple wordlist based on 8 text files from the play Romeo and Juliet, so press *Make a word list now.*
The WordList tool shows us a frequency listing. The most frequent word is "#". There are 985 of these #. Whatever has happened? Well, by default # is used to represent any number such as 65, 40 or $997.82. In this case we have line numbers in the source text.

Below #, the most frequent words are the, and, I to, of. Beside each one you can see how frequent it is in the collection of 8 texts we used, the percentage of running words, and how many of our 8 texts each word occurred in. It seems I is a top frequency word but even so was not present in all the 8 texts.

To see the words in alphabetical order instead, click the alphabetical tab near the bottom of the window.
Now scroll down to *wherefore*. The results seem to confirm what we found when we made a concordance.

### 5.3 concordancing selected words

Once you have a word list on screen, you might want to see some of the words in it in their contexts.

Select a word (or more)
and choose Compute | Concordance.

You will get something like this (if the original texts are still where they were when the word list was first made):

![Concordance screenshot](image)
5.4 lemmatising

To lemmatise manually, with a word list on screen,

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Freq</th>
<th>%</th>
<th>Texts</th>
<th>% lemmas</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,052</td>
<td>BLIN</td>
<td>2</td>
<td>1</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,053</td>
<td>BLIND</td>
<td>111</td>
<td>53</td>
<td>11.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,054</td>
<td>BLINDED</td>
<td>3</td>
<td>3</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,055</td>
<td>BLINDFOLD</td>
<td>3</td>
<td>3</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,056</td>
<td>BLINDFOLDED</td>
<td>4</td>
<td>4</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,057</td>
<td>BLINDING</td>
<td>5</td>
<td>5</td>
<td>1.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,058</td>
<td>BLINDINGLY</td>
<td>2</td>
<td>2</td>
<td>0.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,059</td>
<td>BLINDLY</td>
<td>3</td>
<td>3</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,060</td>
<td>BLINDNESS</td>
<td>8</td>
<td>8</td>
<td>1.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,061</td>
<td>BLINDS</td>
<td>20</td>
<td>4</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,062</td>
<td>BLINDSIDE</td>
<td>1</td>
<td>1</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,063</td>
<td>BLINDSRE</td>
<td>1</td>
<td>1</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,064</td>
<td>BLINIS</td>
<td>1</td>
<td>1</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

pull it onto the line you want to join it to.

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Freq</th>
<th>%</th>
<th>Texts</th>
<th>% lemmas</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,052</td>
<td>BLIN</td>
<td>2</td>
<td>1</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,053</td>
<td>BLIND</td>
<td>111</td>
<td>53</td>
<td>11.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,054</td>
<td>BLINDED</td>
<td>3</td>
<td>3</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,055</td>
<td>BLINDFOLD</td>
<td>3</td>
<td>3</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,056</td>
<td>BLINDFOLDED</td>
<td>4</td>
<td>4</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,057</td>
<td>BLINDING</td>
<td>5</td>
<td>5</td>
<td>1.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,058</td>
<td>BLINDINGLY</td>
<td>2</td>
<td>2</td>
<td>0.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,059</td>
<td>BLINDLY</td>
<td>3</td>
<td>3</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,060</td>
<td>BLINDNESS</td>
<td>8</td>
<td>8</td>
<td>1.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,061</td>
<td>BLINDS</td>
<td>20</td>
<td>4</td>
<td>0.83</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,062</td>
<td>BLINDSIDE</td>
<td>1</td>
<td>1</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,063</td>
<td>BLINDSRE</td>
<td>1</td>
<td>1</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6,064</td>
<td>BLINIS</td>
<td>1</td>
<td>1</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

and drop it:
You will then see the totals change and the items become visible in the Lemmas column.

If there are a lot, you can double-click the Lemmas column to see the details:
5.5 word list statistics

Press the statistics tab at the bottom of a word list,

and something like this should appear. Lots of numbers. Further down, the numbers are easier to understand:
There are lots of 4-letter words in Shakespeare, it seems.

5.6 multi-word units

5.6.1 using an index

To make a wordlist with pairs or triples of words (n-grams) such as

    OF THE
    IN THE END
    ONCE UPON A TIME

etc you will need first to compute an index file. This essentially knows the position of each separate word in your corpus.

See also : making the multi-word unit wordlist
5.6.2 making a multi-word wordlist

The process is explained here and what you get looks like this.

Press Ctrl/F2 to save it, and the suggested filename will be something like `_index_3-5-word clusters`. It can later be opened as an ordinary wordlist.
Step-by-step guide to WordSmith

KeyWords

Section VI
6 KeyWords

6.1 overview

A key word list in WordSmith Tools looks something like this.

![KeyWords window](image)

The key words are words which occur unusually frequently in comparison with some kind of reference corpus.

Beside each key word there are various numbers telling you how frequent each one was in the source text(s) and how that compares with its frequency in the reference corpus.

In the list above, based on the play Romeo and Juliet in comparison with all the Shakespeare plays, we see lots of names of the main characters, some pronouns like *thou*, plus theme words like *love* and *night*.

6.2 making a key word list

To make a key word list, first press the KeyWords button in the main Controller.
When KeyWords starts up, choose menu option File, then New and you will see something like this.

You have to choose word lists made and saved by WordSmith Tools.

You can choose the word list files by pressing this button:

The reference corpus word list is assumed to be a big one, which will help WordSmith work out what is unusual about the words in your chosen text(s).

Once you have chosen a word list above and another for your reference below, press Make a keyword list now. (Until you have, that button won’t be enabled.) Then you will see something like this:
6.3  **key words plot**

This is a key word plot where the text is the file `a1f` in the British National Corpus (BNC), compared with the whole of the BNC.
You see:

- each key word (KW) (these obviously have to do with international relations)
- a measure of its dispersion and its keyness
- how many times each KW came in the text (hits).
- a map showing where each word came.

At the left the blue line represents the start of the text, at the right the blue line represents the end. Look at Britain, Germany, Italy and century -- these seem to come in bursts more or less three-quarters of the way through the text. China, Mao, Peking come together a bit later in the text.

6.4 concordancing selected key words

Once you have a key word list on screen, you might want to see some of the words in it in their contexts.
Select a word (or more)
and choose **Compute | Concordance.** Here, the rather mysterious HAH has been chosen.

You will get something like this (if the original texts are still where they were when the word list was first made):

It seems that HAH is a system for health care.
Index

- C -
  choosing text files  4
  collocates and mutual information  17
  Concord: nearest tag  21
  Concord: overview  12
  concordancing on tags  20

- I -
  introduction  2

- K -
  KeyWords: overview  35

- M -
  making a concordance  13

- N -
  nearest tag  21

- S -
  seeing source text  15
  sorting tags  21

- T -
  tag concordancing  20

- W -
  WordList: overview  24