WordSmith Tools step by step

Version 5.0

WordSmith Tools step by step

version 5.0

by Mike Scott

Table of Contents

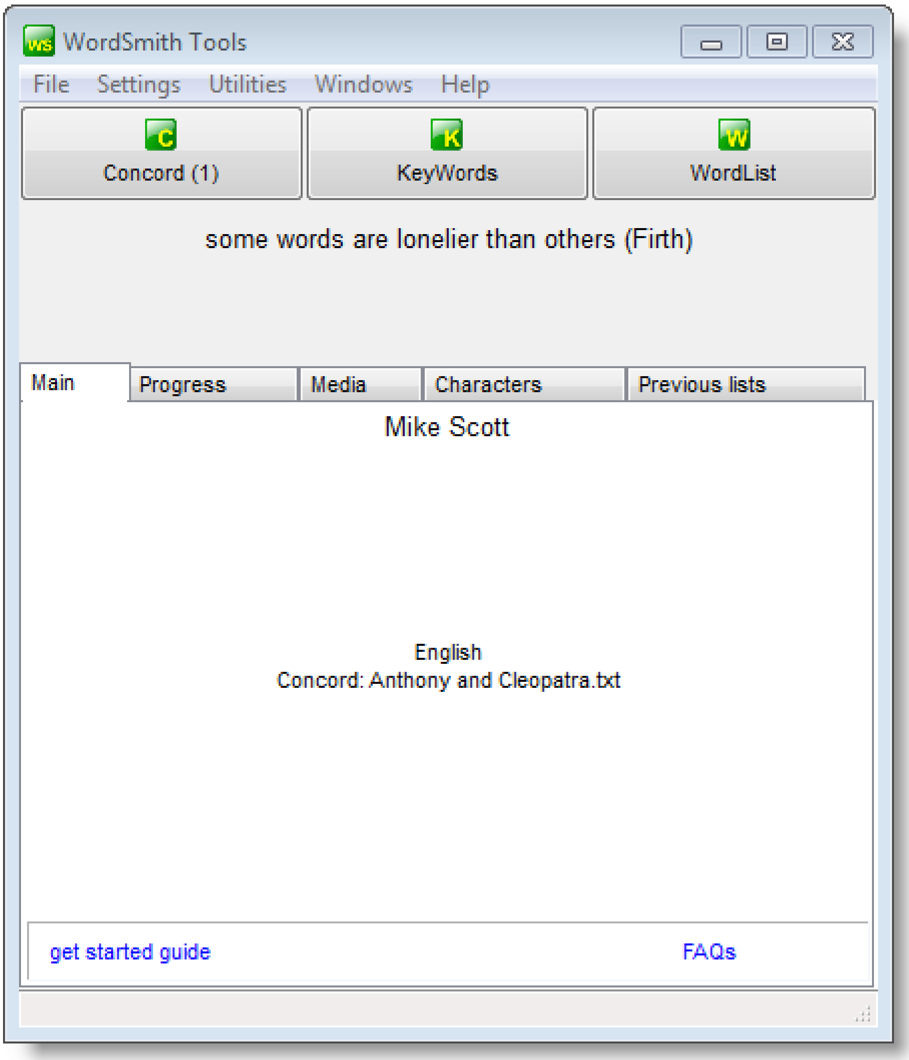
Introduction

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.



some words are lonelier than others (Firth)

Mike Scott

English
Concord: Anthony and Cleopatra.txt

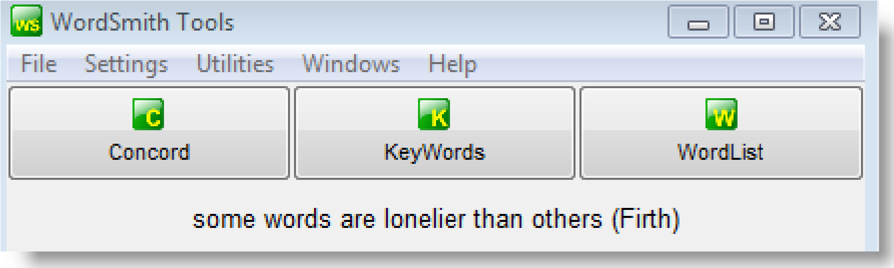
It has a saying (which keeps on changing and which you can edit), three buttons for the main Tools (Concord is shown as in use), and a series of tabs. At the moment we see the main one showing that **anthony & cleopatra.txt** has been chosen for Concord.

Choosing your texts

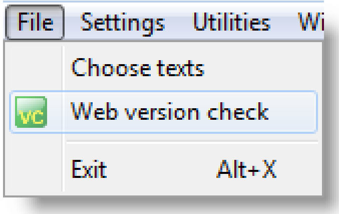
Section II

2 Choosing your texts

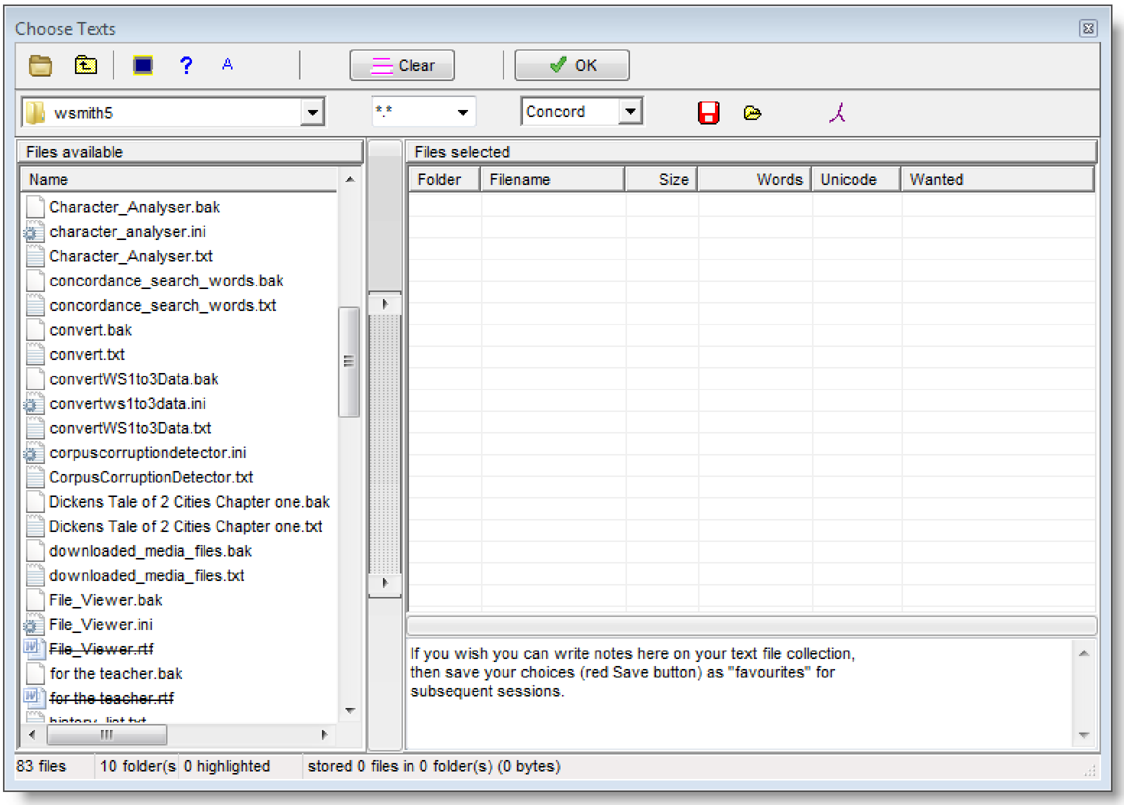
To choose text files, click the File menu in the main Controller:



some words are lonelier than others (Firth)



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



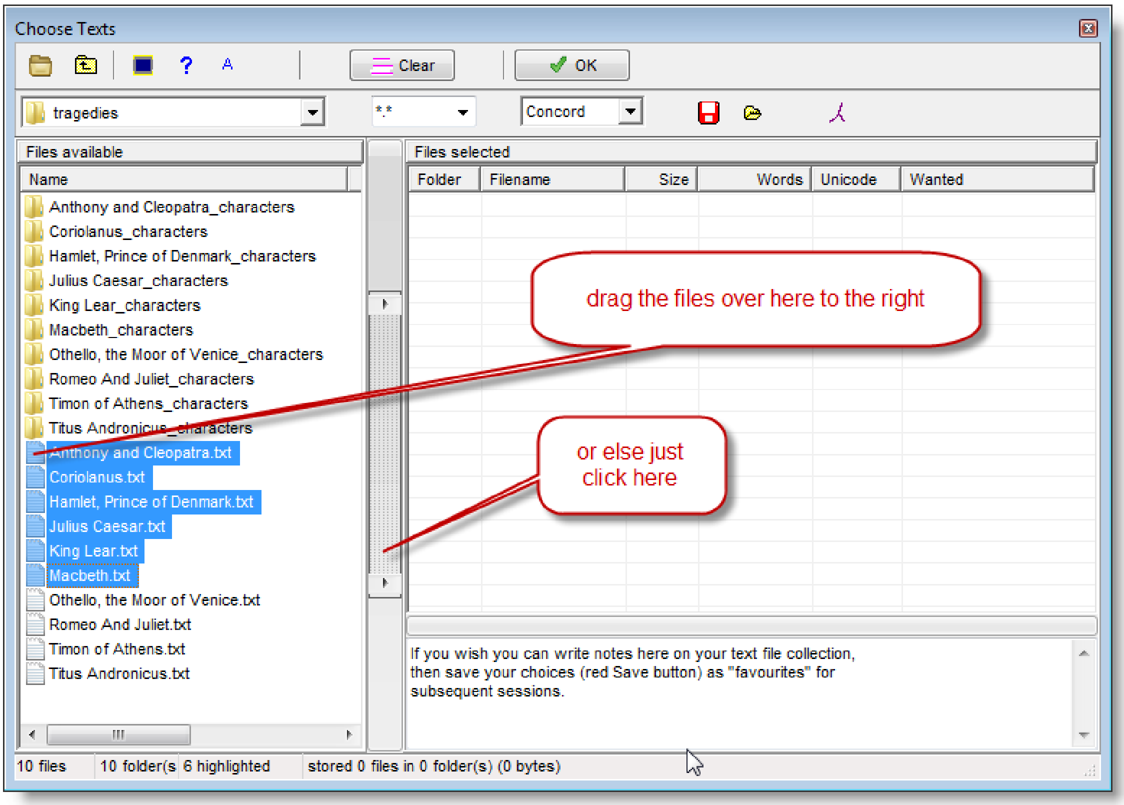
Name
Character_Analyser.bak
character_analyser.ini
Character_Analyser.txt
concordance_search_words.bak
concordance_search_words.txt
convert.bak
convert.txt
convertWS1to3Data.bak
convertws1to3data.ini
convertWS1to3Data.txt
corpuscorruptiondetector.ini
CorpusCorruptionDetector.txt
Dickens Tale of 2 Cities Chapter one.bak
Dickens Tale of 2 Cities Chapter one.txt
downloaded_media_files.bak
downloaded_media_files.txt
File_Viewer.bak
File_Viewer.ini
~~File_Viewer.rtf~~
for the teacher.bak
~~for the teacher.rtf~~
history_list.txt

If you wish you can write notes here on your text file collection, then save your choices (red Save button) as "favourites" for subsequent sessions.

83 files | 10 folder(s | 0 highlighted | stored 0 files in 0 folder(s) (0 bytes)

At the left is a fairly standard text file explorer, at the right an area for Files selected.

Press the browse button (Choose) to find the folder where your texts are. You need plain text (. `txt`) files.



Anthony and Cleopatra_characters
Coriolanus_characters
Hamlet, Prince of Denmark_characters
Julius Caesar_characters
King Lear_characters
Macbeth_characters
Othello, the Moor of Venice_characters
Romeo And Juliet_characters
Timon of Athens_characters
Titus Andronicus_characters
Anthony and Cleopatra.txt
Coriolanus.txt
Hamlet, Prince of Denmark.txt
Julius Caesar.txt
King Lear.txt
Macbeth.txt
Othello, the Moor of Venice.txt
Romeo And Juliet.txt
Timon of Athens.txt
Titus Andronicus.txt

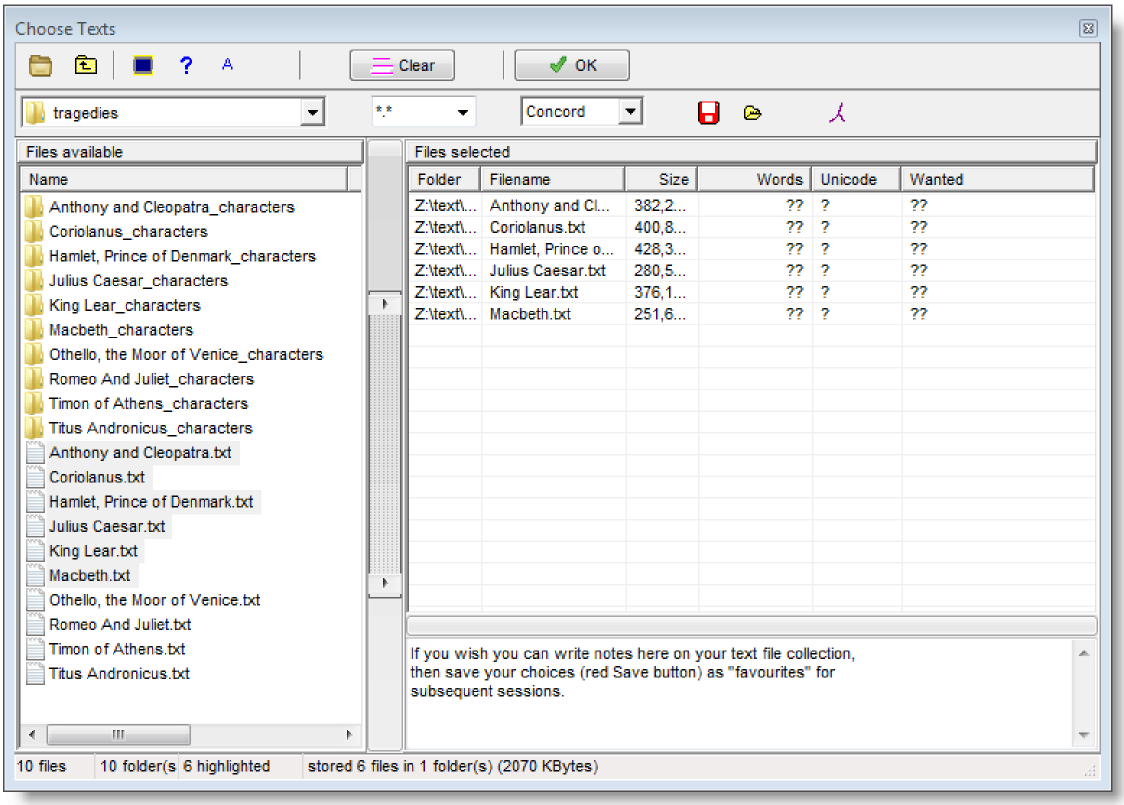
drag the files over here to the right

or else just click here

If you wish you can write notes here on your text file collection, then save your choices (red Save button) as "favourites" for subsequent sessions.

10 files | 10 folder(s 6 highlighted | stored 0 files in 0 folder(s) (0 bytes)

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



Files available

Name
Anthony and Cleopatra_characters
Coriolanus_characters
Hamlet, Prince of Denmark_characters
Julius Caesar_characters
King Lear_characters
Macbeth_characters
Othello, the Moor of Venice_characters
Romeo And Juliet_characters
Timon of Athens_characters
Titus Andronicus_characters
Anthony and Cleopatra.txt
Coriolanus.txt
Hamlet, Prince of Denmark.txt
Julius Caesar.txt
King Lear.txt
Macbeth.txt
Othello, the Moor of Venice.txt
Romeo And Juliet.txt
Timon of Athens.txt
Titus Andronicus.txt

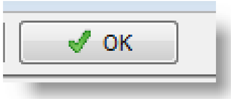
Files selected

Folder	Filename	Size	Words	Unicode	Wanted
Z:\text\...	Anthony and Cl...	382,2...	??	?	??
Z:\text\...	Coriolanus.txt	400,8...	??	?	??
Z:\text\...	Hamlet, Prince o...	428,3...	??	?	??
Z:\text\...	Julius Caesar.txt	280,5...	??	?	??
Z:\text\...	King Lear.txt	376,1...	??	?	??
Z:\text\...	Macbeth.txt	251,6...	??	?	??

If you wish you can write notes here on your text file collection, then save your choices (red Save button) as "favourites" for subsequent sessions.

10 files | 10 folder(s 6 highlighted | stored 6 files in 1 folder(s) (2070 KBytes)

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*).



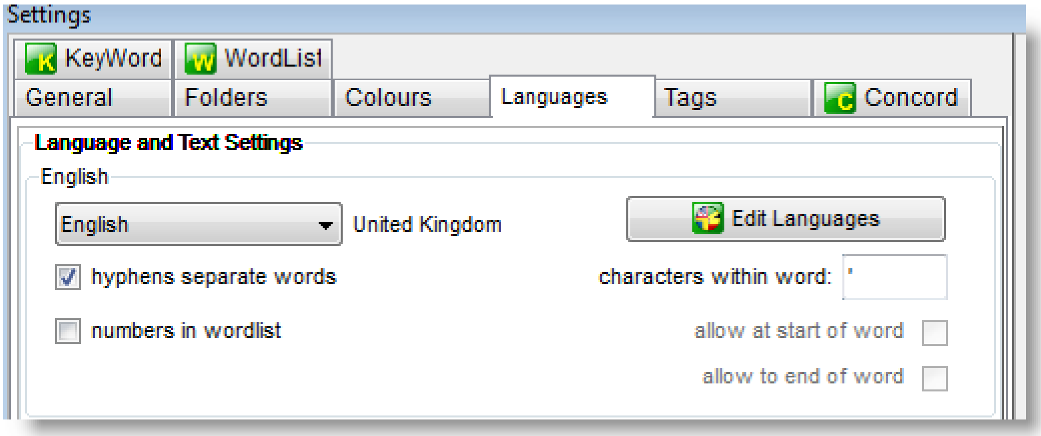
Press the green button or just close the window.

Select the right language

Section III

3 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.



Language and Text Settings

English

English United Kingdom

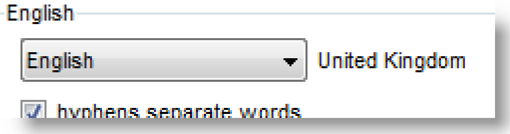
☑ hyphens separate words

☐ numbers in wordlist

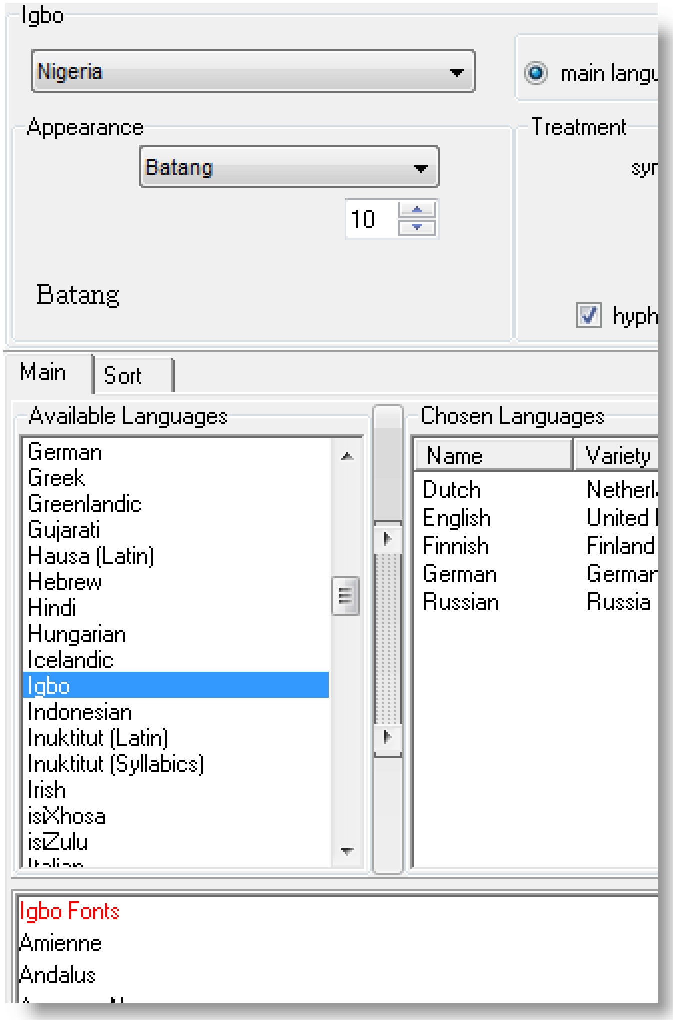
characters within word: '

allow at start of word ☐

allow to end of word ☐



If in the drop-down list (showing) the language you need isn't available, click *Edit Languages* and choose the language you want*:*



Igbo

Nigeria

● main langu

Appearance

Batang

10

Batang

Treatment

syn

☑ hyph

Available Languages

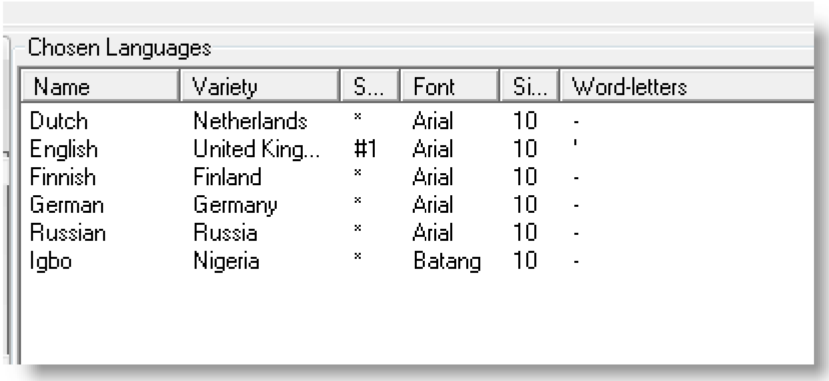
German
Greek
Greenlandic
Gujarati
Hausa (Latin)
Hebrew
Hindi
Hungarian
Icelandic
Igbo
Indonesian
Inuktitut (Latin)
Inuktitut (Syllabics)
Irish
isiXhosa
isiZulu
Italian

Chosen Languages

Name	Variety
Dutch	Netherl
English	United I
Finnish	Finland
German	German
Russian	Russia

Igbo Fonts
Amienne
Andalus

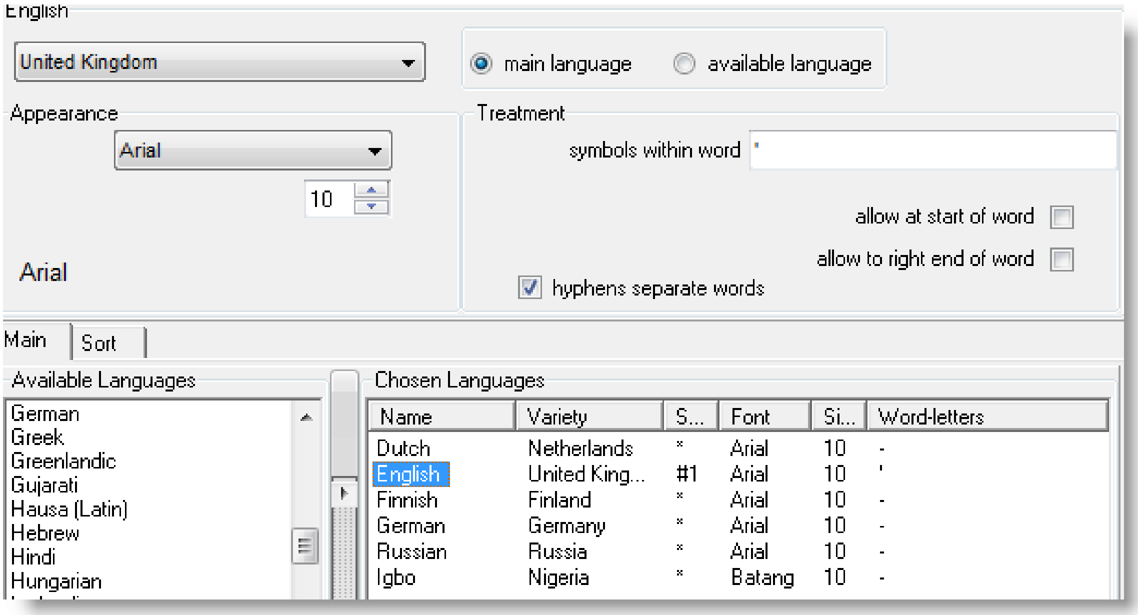
and drag it to the right or click the button in the middle.



Chosen Languages

Name	Variety	S...	Font	Si...	Word-letters
Dutch	Netherlands	*	Arial	10	-
English	United King...	#1	Arial	10	'
Finnish	Finland	*	Arial	10	-
German	Germany	*	Arial	10	-
Russian	Russia	*	Arial	10	-
Igbo	Nigeria	*	Batang	10	-

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.


English

United Kingdom

● main language ○ available language

Appearance

Arial

10

Arial

Treatment

symbols within word '

allow at start of word □

allow to right end of word □

☑ hyphens separate words

Available Languages

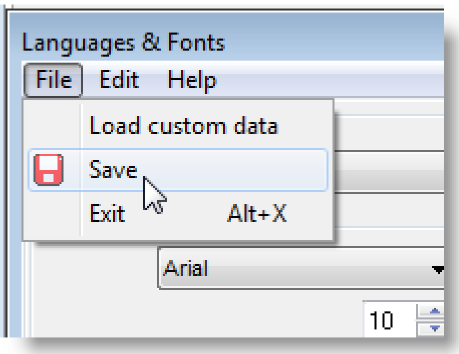
German
Greek
Greenlandic
Gujarati
Hausa (Latin)
Hebrew
Hindi
Hungarian

Chosen Languages

Name	Variety	S...	Font	Si...	Word-letters
Dutch	Netherlands	*	Arial	10	-
English	United King...	#1	Arial	10	'
Finnish	Finland	*	Arial	10	-
German	Germany	*	Arial	10	-
Russian	Russia	*	Arial	10	-
Igbo	Nigeria	*	Batang	10	-

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.



Step-by-step guide to WordSmith

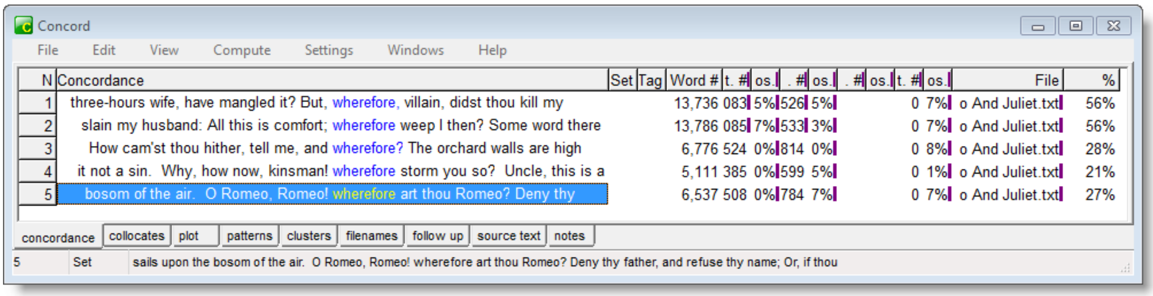
Concordancing

Section IV

4 Concordancing

4.1 overview

A concordance looks something like this:



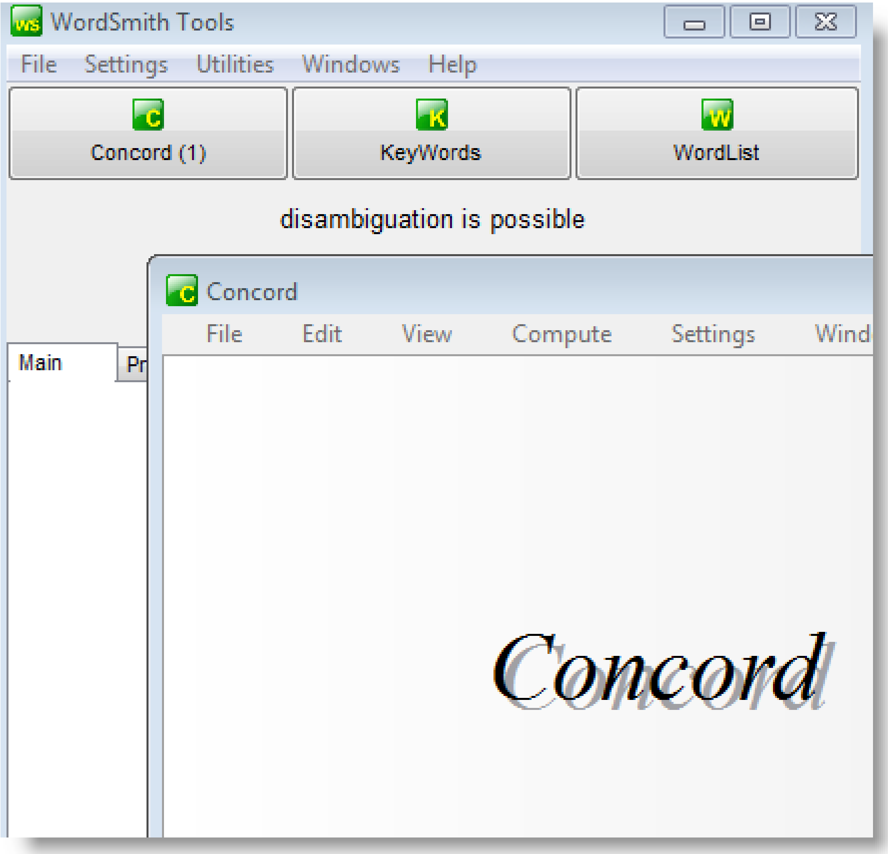
N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
1	three-hours wife, have mangled it? But, wherefore, villain, didst thou kill my			13,736	083	5%	526	5%			0	7%	o And Juliet.txt	56%
2	slain my husband: All this is comfort; wherefore weep I then? Some word there			13,786	085	7%	533	3%			0	7%	o And Juliet.txt	56%
3	How cam'st thou hither, tell me, and wherefore? The orchard walls are high			6,776	524	0%	814	0%			0	8%	o And Juliet.txt	28%
4	it not a sin. Why, how now, kinsman! wherefore storm you so? Uncle, this is a			5,111	385	0%	599	5%			0	1%	o And Juliet.txt	21%
5	bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy			6,537	508	0%	784	7%			0	7%	o And Juliet.txt	27%

5 Set sails upon the bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy father, and refuse thy name; Or, if thou

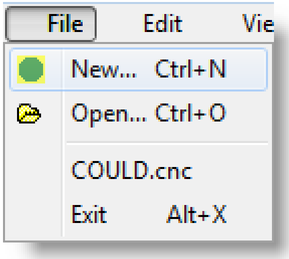
It's a concordance of all the occurrences of **wherefore** in Romeo and Juliet. There are only 5 entries. The famous one comes 6,537 words (27%) into the play.

4.2 making a concordance

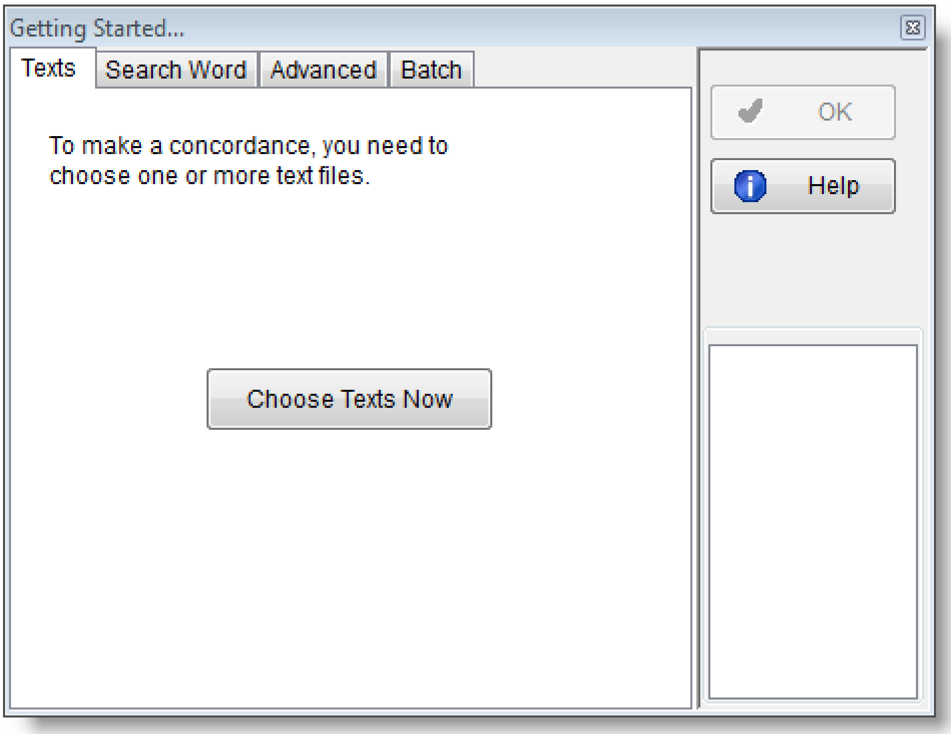
When you press the Concord button in the main Controller, a new Concord Tool opens up and will be visible in the Windows Taskbar.


WordSmith Tools
File Settings Utilities Windows Help
Concord (1)
KeyWords
WordList
disambiguation is possible
Concord
File Edit View Compute Settings Wind
Main
Pr
Concord

Now in Concord itself, choose *File* | *New*.

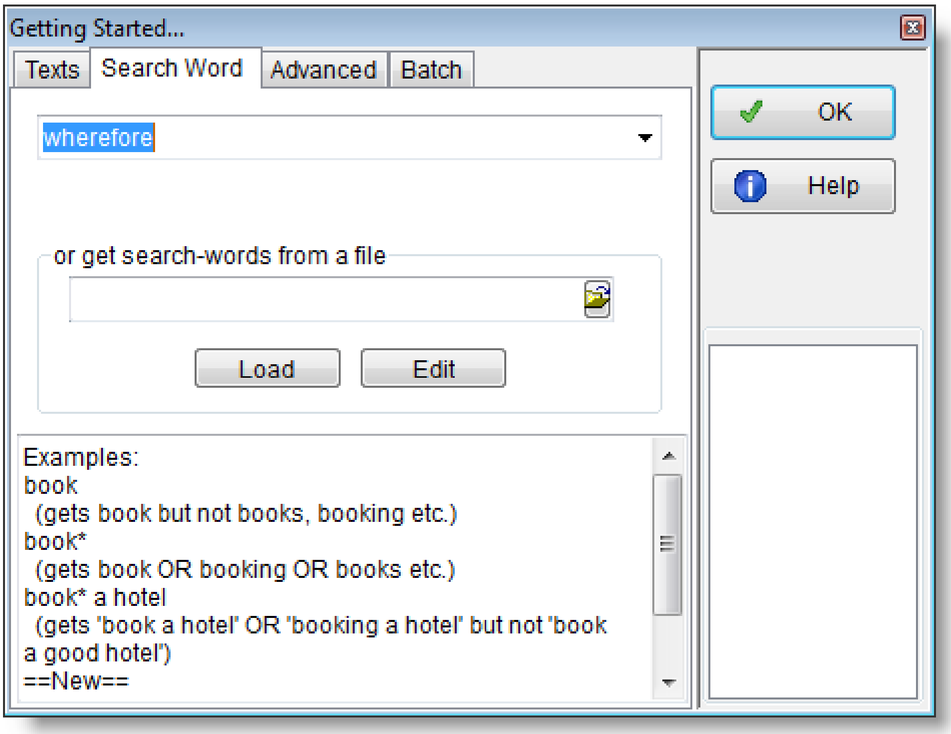


If no text files have been chosen, you are asked to choose some. Press the *Choose Texts Now* button.



To make a concordance, you need to choose one or more text files.

Once the texts have been chosen, enter a suitable *Search Word*:

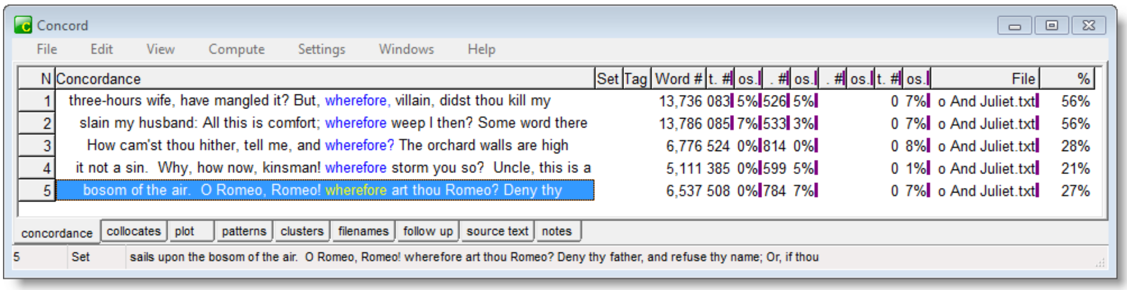


wherefore

or get search-words from a file

Examples:
book
(gets book but not books, booking etc.)
book*
(gets book OR booking OR books etc.)
book* a hotel
(gets 'book a hotel' OR 'booking a hotel' but not 'book a good hotel')
==New==

Here I have chosen **`wherefore`** as my search-word. Then press OK.

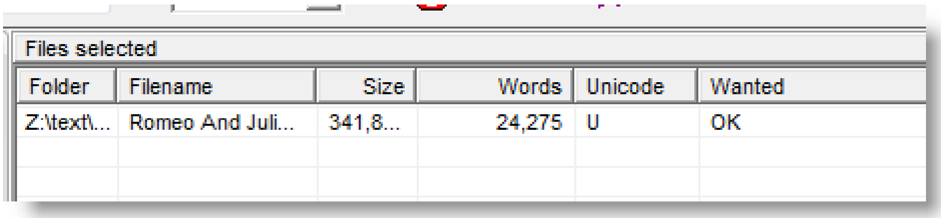


N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
1	three-hours wife, have mangled it? But, wherefore, villain, didst thou kill my			13,736	083	5%	526	5%			0	7%	o And Juliet.txt	56%
2	slain my husband: All this is comfort; wherefore weep I then? Some word there			13,786	085	7%	533	3%			0	7%	o And Juliet.txt	56%
3	How cam'st thou hither, tell me, and wherefore? The orchard walls are high			6,776	524	0%	814	0%			0	8%	o And Juliet.txt	28%
4	it not a sin. Why, how now, kinsman! wherefore storm you so? Uncle, this is a			5,111	385	0%	599	5%			0	1%	o And Juliet.txt	21%
5	bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy			6,537	508	0%	784	7%			0	7%	o And Juliet.txt	27%

5 Set sails upon the bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy father, and refuse thy name; Or, if thou

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.

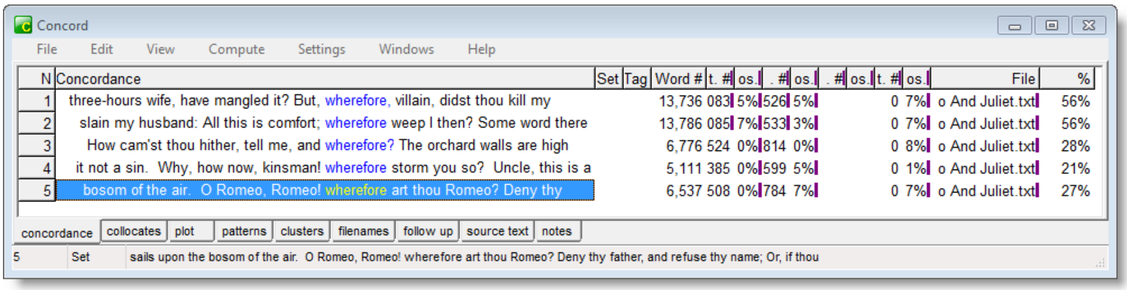


Files selected

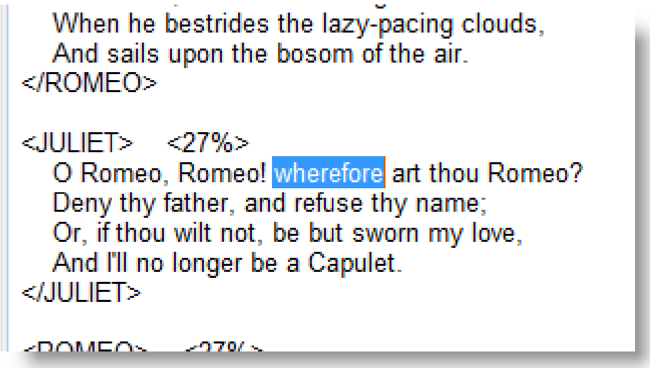
Folder	Filename	Size	Words	Unicode	Wanted
Z:\text\...	Romeo And Juli...	341,8...	24,275	U	OK

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*.

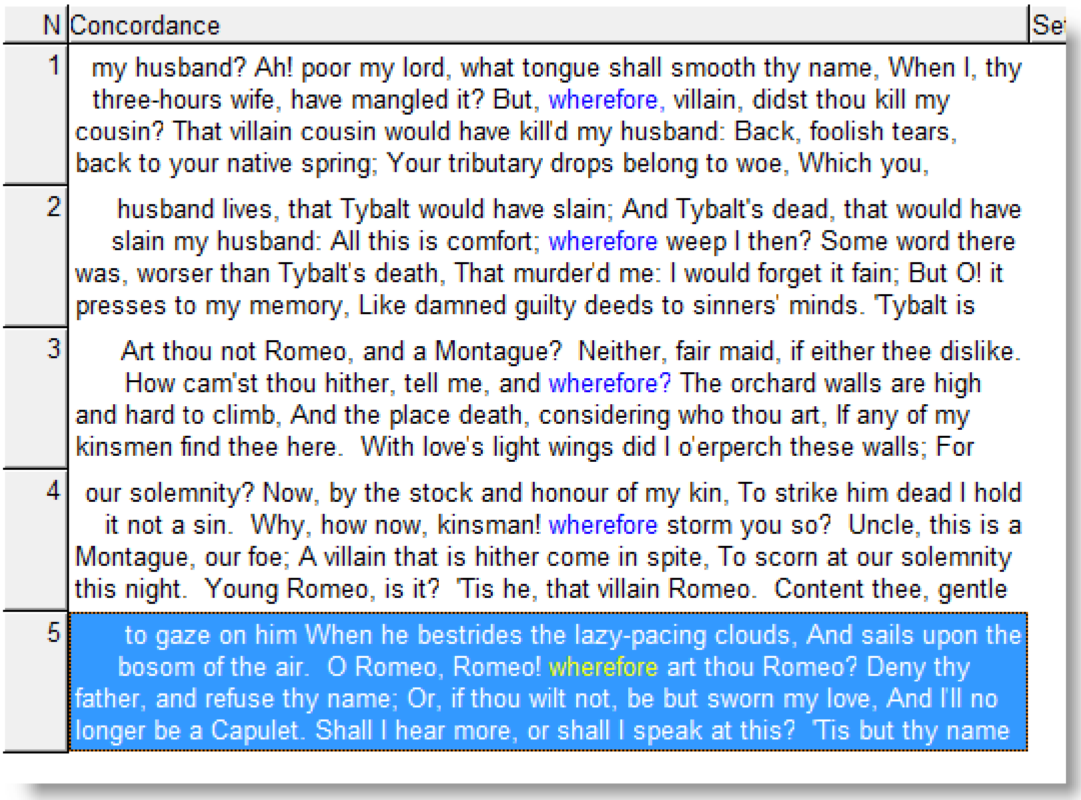

N	Concordance	Set	Tag	Word #	t. #	os.	. #	os.	. #	os.	t. #	os.	File	%
1	three-hours wife, have mangled it? But, wherefore, villain, didst thou kill my			13,736	083	5%	526	5%			0	7%	o And Juliet.txt	56%
2	slain my husband: All this is comfort; wherefore weep I then? Some word there			13,786	085	7%	533	3%			0	7%	o And Juliet.txt	56%
3	How cam'st thou hither, tell me, and wherefore? The orchard walls are high			6,776	524	0%	814	0%			0	8%	o And Juliet.txt	28%
4	it not a sin. Why, how now, kinsman! wherefore storm you so? Uncle, this is a			5,111	385	0%	599	5%			0	1%	o And Juliet.txt	21%
5	bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy			6,537	508	0%	784	7%			0	7%	o And Juliet.txt	27%

5 Set sails upon the bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy father, and refuse thy name; Or, if thou



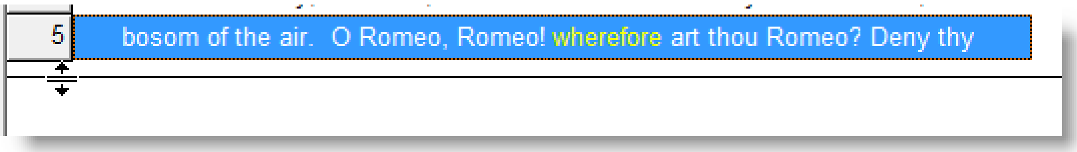
```
   When he bestrides the lazy-pacing clouds,
   And sails upon the bosom of the air.
</ROMEO>

<JULIET>    <27%>
   O Romeo, Romeo! wherefore art thou Romeo?
   Deny thy father, and refuse thy name;
   Or, if thou wilt not, be but sworn my love,
   And I'll no longer be a Capulet.
</JULIET>

<ROMEO>    <27%>
```

or press F8 and the lines grow fatter:



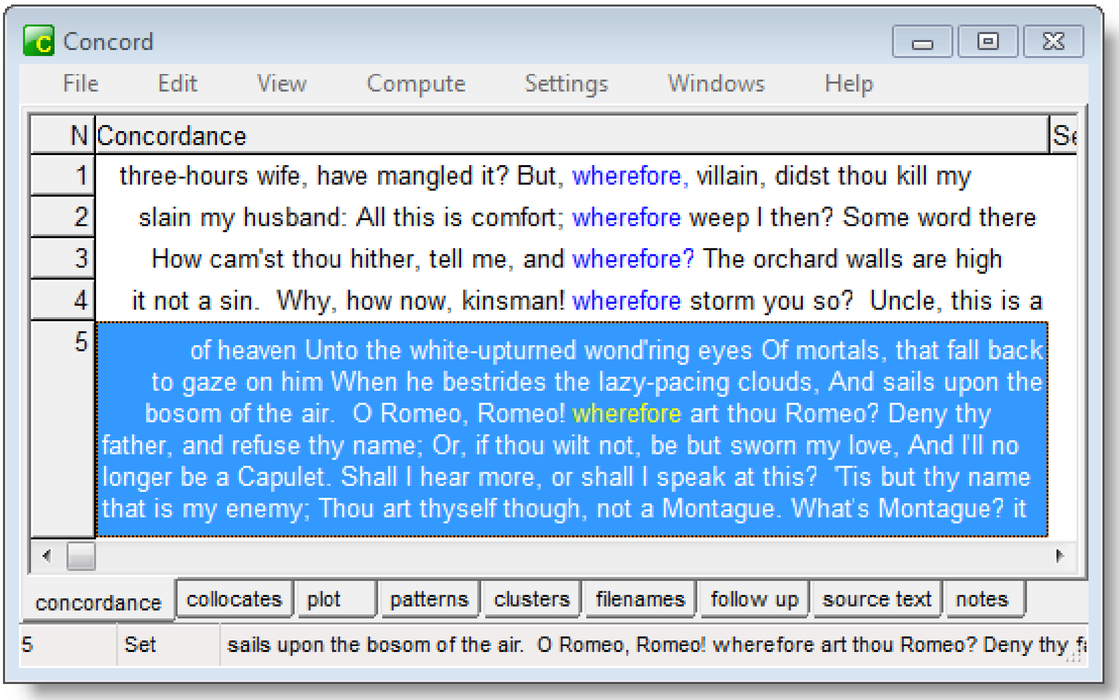
N	Concordance	Se
1	my husband? Ah! poor my lord, what tongue shall smooth thy name, When I, thy three-hours wife, have mangled it? But, wherefore, villain, didst thou kill my cousin? That villain cousin would have kill'd my husband: Back, foolish tears, back to your native spring; Your tributary drops belong to woe, Which you,	
2	husband lives, that Tybalt would have slain; And Tybalt's dead, that would have slain my husband: All this is comfort; wherefore weep I then? Some word there was, worser than Tybalt's death, That murder'd me: I would forget it fain; But O! it presses to my memory, Like damned guilty deeds to sinners' minds. Tybalt is	
3	Art thou not Romeo, and a Montague? Neither, fair maid, if either thee dislike. How cam'st thou hither, tell me, and wherefore? The orchard walls are high and hard to climb, And the place death, considering who thou art, If any of my kinsmen find thee here. With love's light wings did I o'erperch these walls; For	
4	our solemnity? Now, by the stock and honour of my kin, To strike him dead I hold it not a sin. Why, how now, kinsman! wherefore storm you so? Uncle, this is a Montague, our foe; A villain that is hither come in spite, To scorn at our solemnity this night. Young Romeo, is it? 'Tis he, that villain Romeo. Content thee, gentle	
5	to gaze on him When he bestrides the lazy-pacing clouds, And sails upon the bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy father, and refuse thy name; Or, if thou wilt not, be but sworn my love, And I'll no longer be a Capulet. Shall I hear more, or shall I speak at this? 'Tis but thy name	

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.



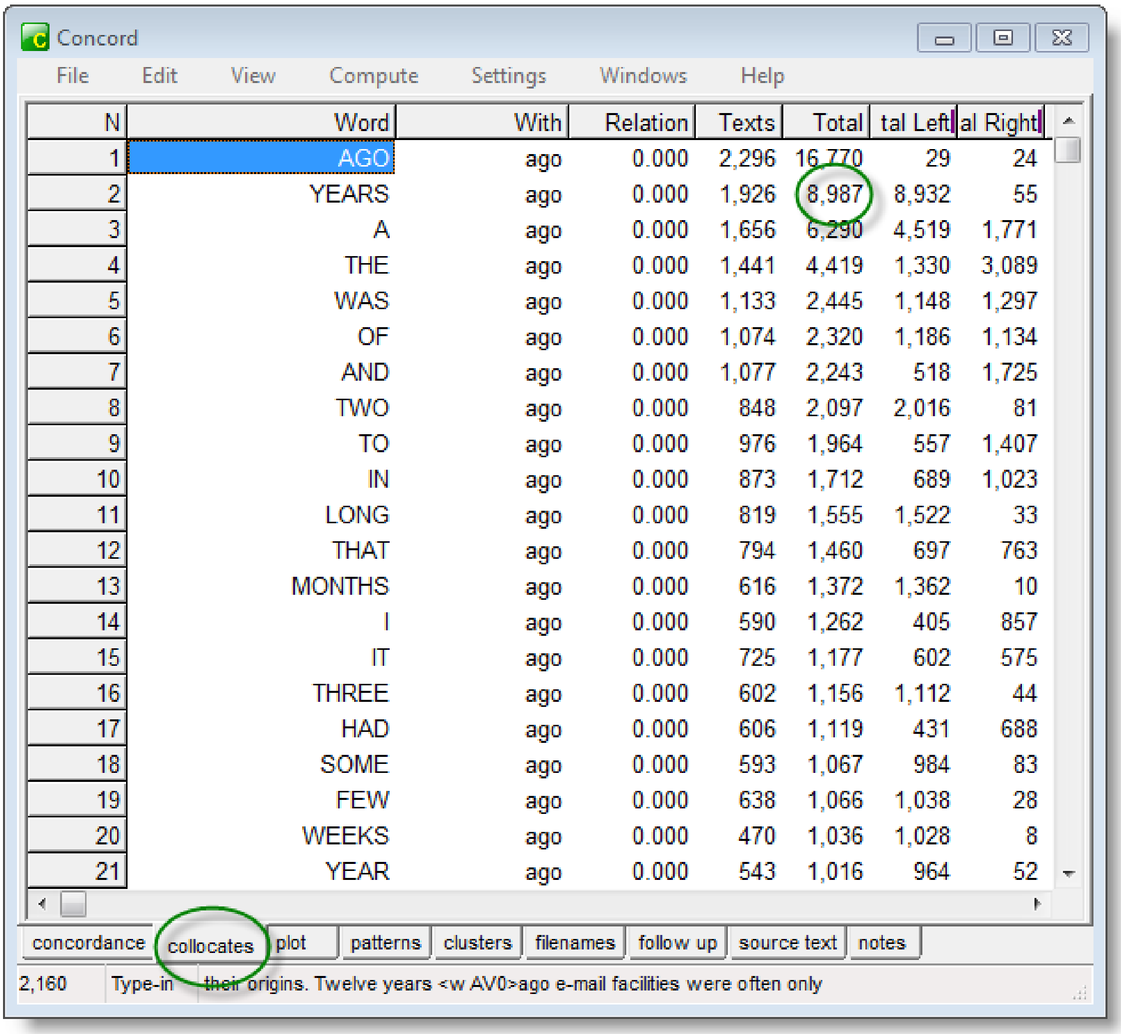
N	Concordance	Se
1	three-hours wife, have mangled it? But, wherefore, villain, didst thou kill my	
2	slain my husband: All this is comfort; wherefore weep I then? Some word there	
3	How cam'st thou hither, tell me, and wherefore? The orchard walls are high	
4	it not a sin. Why, how now, kinsman! wherefore storm you so? Uncle, this is a	
5	of heaven Unto the white-upturned wond'ring eyes Of mortals, that fall back to gaze on him When he bestrides the lazy-pacing clouds, And sails upon the bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy father, and refuse thy name; Or, if thou wilt not, be but sworn my love, And I'll no longer be a Capulet. Shall I hear more, or shall I speak at this? 'Tis but thy name that is my enemy; Thou art thyself though, not a Montague. What's Montague? it	

5 Set sails upon the bosom of the air. O Romeo, Romeo! wherefore art thou Romeo? Deny thy f

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.

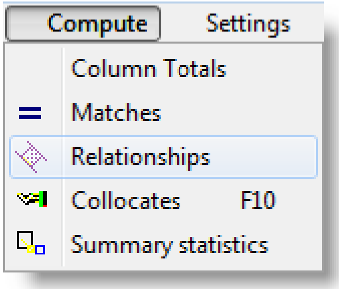


N	Word	With	Relation	Texts	Total	tal Left	al Right
1	AGO	ago	0.000	2,296	16,770	29	24
2	YEARS	ago	0.000	1,926	8,987	8,932	55
3	A	ago	0.000	1,656	6,290	4,519	1,771
4	THE	ago	0.000	1,441	4,419	1,330	3,089
5	WAS	ago	0.000	1,133	2,445	1,148	1,297
6	OF	ago	0.000	1,074	2,320	1,186	1,134
7	AND	ago	0.000	1,077	2,243	518	1,725
8	TWO	ago	0.000	848	2,097	2,016	81
9	TO	ago	0.000	976	1,964	557	1,407
10	IN	ago	0.000	873	1,712	689	1,023
11	LONG	ago	0.000	819	1,555	1,522	33
12	THAT	ago	0.000	794	1,460	697	763
13	MONTHS	ago	0.000	616	1,372	1,362	10
14	I	ago	0.000	590	1,262	405	857
15	IT	ago	0.000	725	1,177	602	575
16	THREE	ago	0.000	602	1,156	1,112	44
17	HAD	ago	0.000	606	1,119	431	688
18	SOME	ago	0.000	593	1,067	984	83
19	FEW	ago	0.000	638	1,066	1,038	28
20	WEEKS	ago	0.000	470	1,036	1,028	8
21	YEAR	ago	0.000	543	1,016	964	52

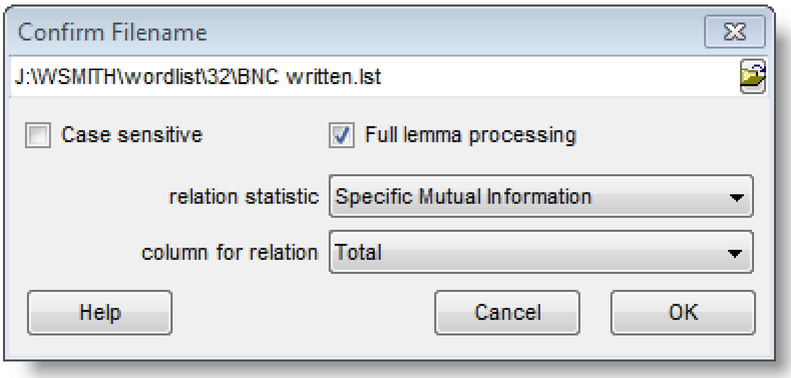
2,160 Type-in their origins. Twelve years <w AV0>ago e-mail facilities were often only

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:



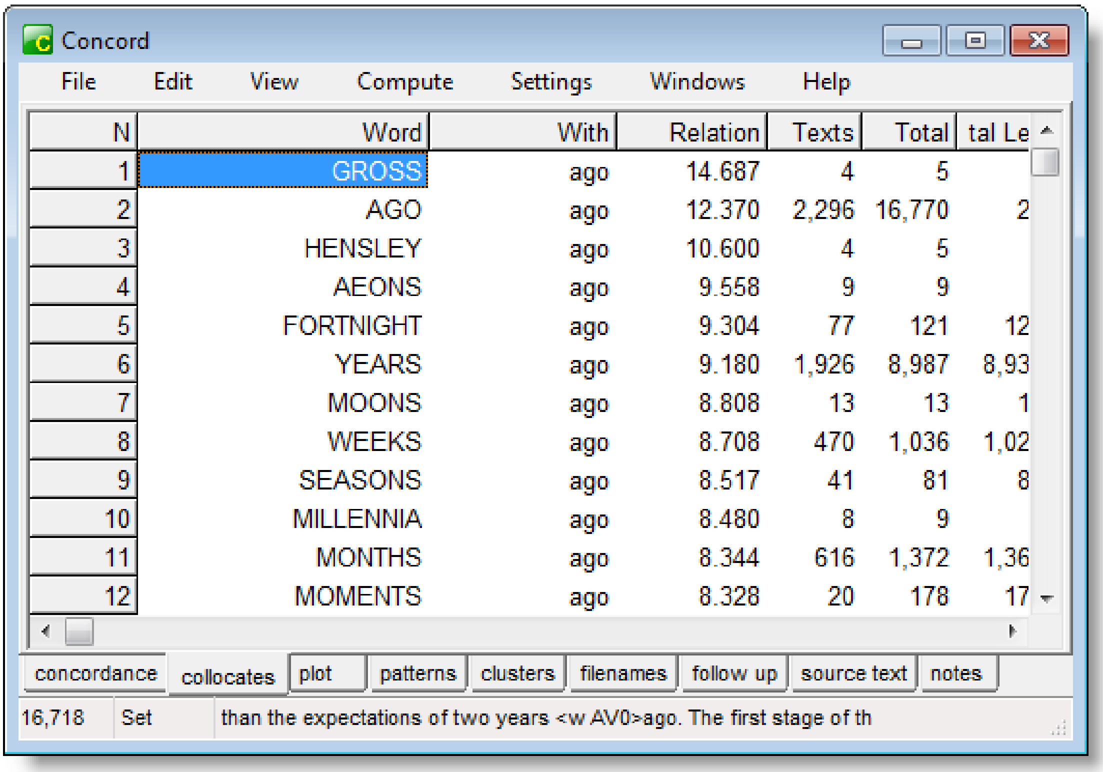
J:\WSMITH\wordlist\32\BNC written.lst

☐ Case sensitive ☑ Full lemma processing

relation statistic: Specific Mutual Information

column for relation: Total

then we get the following list when sorted by clicking the *Relation* column:



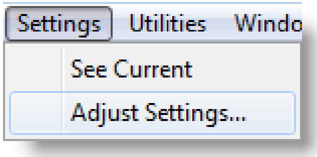
N	Word	With	Relation	Texts	Total	tal Le
1	GROSS	ago	14.687	4	5	
2	AGO	ago	12.370	2,296	16,770	2
3	HENSLEY	ago	10.600	4	5	
4	AEONS	ago	9.558	9	9	
5	FORTNIGHT	ago	9.304	77	121	12
6	YEARS	ago	9.180	1,926	8,987	8,93
7	MOONS	ago	8.808	13	13	1
8	WEEKS	ago	8.708	470	1,036	1,02
9	SEASONS	ago	8.517	41	81	8
10	MILLENNIA	ago	8.480	8	9	
11	MONTHS	ago	8.344	616	1,372	1,36
12	MOMENTS	ago	8.328	20	178	17

16,718 Set than the expectations of two years <w AV0>ago. The first stage of th

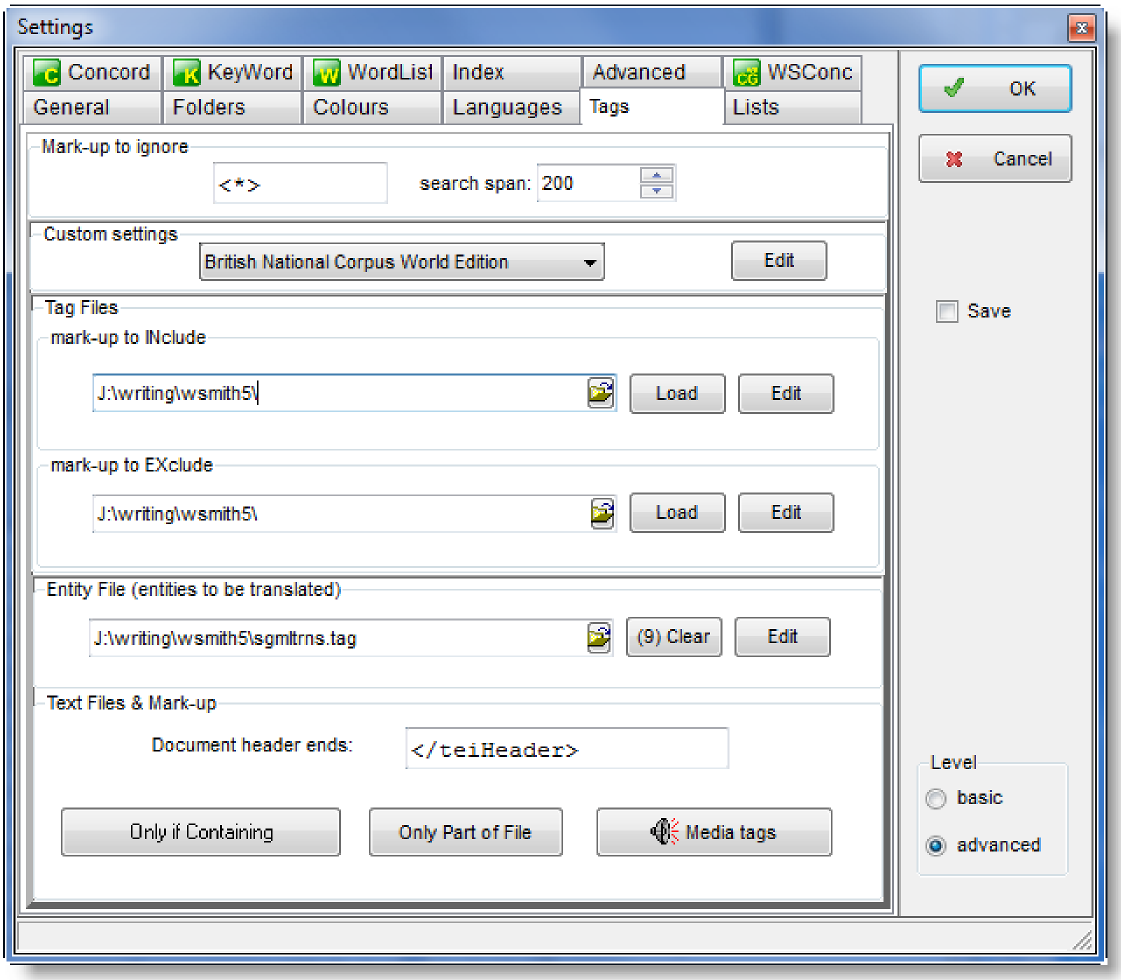
The top items in the list now reflect much better the tendency of **AGO** to accompany periods of time and numbers. [The top collocates **HENSLEY and GROSS** only occur 5 times each with **AGO** but out of small numbers altogether in the whole BNC Written.]

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*.



Mark-up to ignore: <*> search span: 200

Custom settings: British National Corpus World Edition

Tag Files

mark-up to INclude: J:\writing\wsmith5\

mark-up to EXclude: J:\writing\wsmith5\

Entity File (entities to be translated): J:\writing\wsmith5\sgmltrns.tag

Text Files & Mark-up

Document header ends: `</teiHeader>`

□ Save

Level
○ basic
● advanced

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.

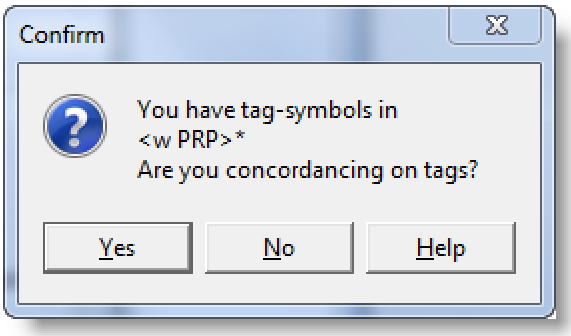
4.6 concordancing tagged text (2)

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

```
<w PRP>at <w AT0>the <w AJ0>great <w NN2>houses
```

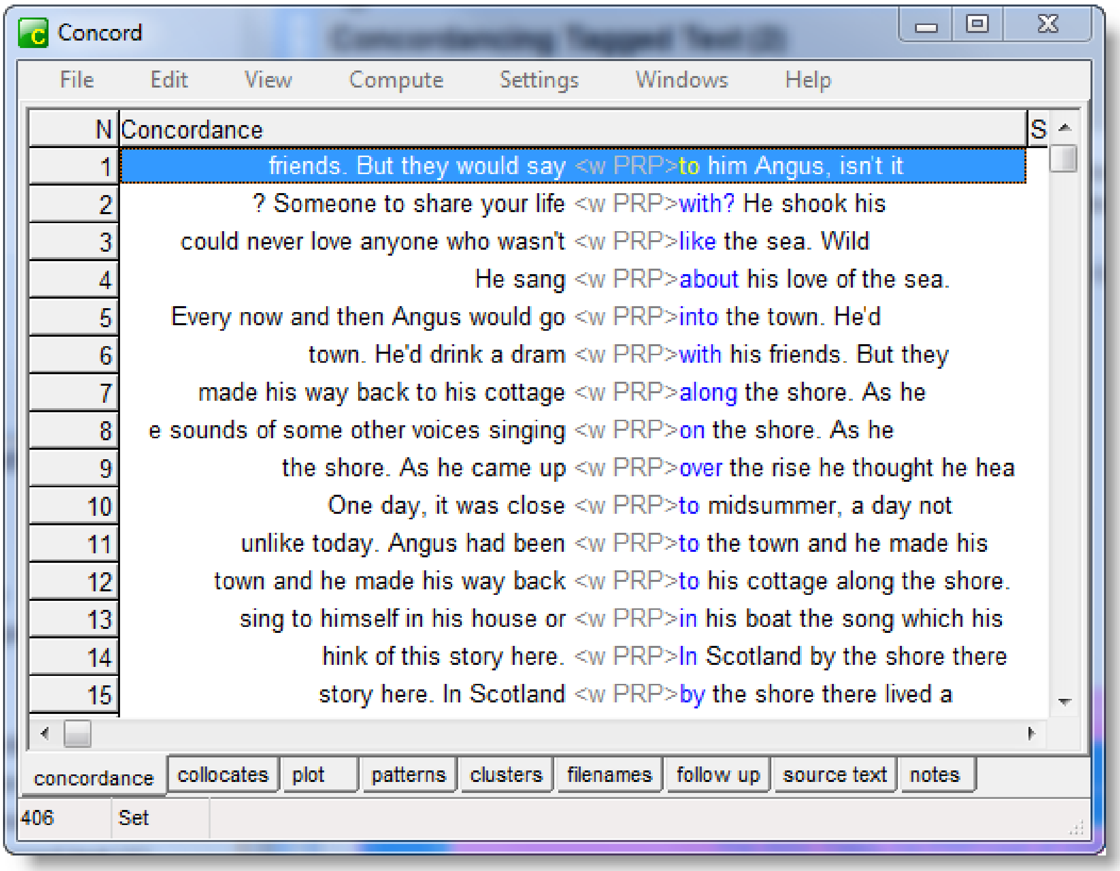
so each preposition is marked **`<w PRP>`** just before the preposition itself. The aim is to see all the prepositions in a selected text from the BNC. With a BNC text file chosen, type **`<w PRP>*`** as the search-word (the asterisk is needed because a word follows directly after the part-of speech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

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



You have tag-symbols in
<w PRP>*
Are you concordancing on tags?

Here we answer Yes.



N	Concordance	S
1	friends. But they would say <w PRP>to him Angus, isn't it	
2	? Someone to share your life <w PRP>with? He shook his	
3	could never love anyone who wasn't <w PRP>like the sea. Wild	
4	He sang <w PRP>about his love of the sea.	
5	Every now and then Angus would go <w PRP>into the town. He'd	
6	town. He'd drink a dram <w PRP>with his friends. But they	
7	made his way back to his cottage <w PRP>along the shore. As he	
8	e sounds of some other voices singing <w PRP>on the shore. As he	
9	the shore. As he came up <w PRP>over the rise he thought he hea	
10	One day, it was close <w PRP>to midsummer, a day not	
11	unlike today. Angus had been <w PRP>to the town and he made his	
12	town and he made his way back <w PRP>to his cottage along the shore.	
13	sing to himself in his house or <w PRP>in his boat the song which his	
14	hink of this story here. <w PRP>In Scotland by the shore there	
15	story here. In Scotland <w PRP>by the shore there lived a	

You see the prepositions and their tags (but no other tags).

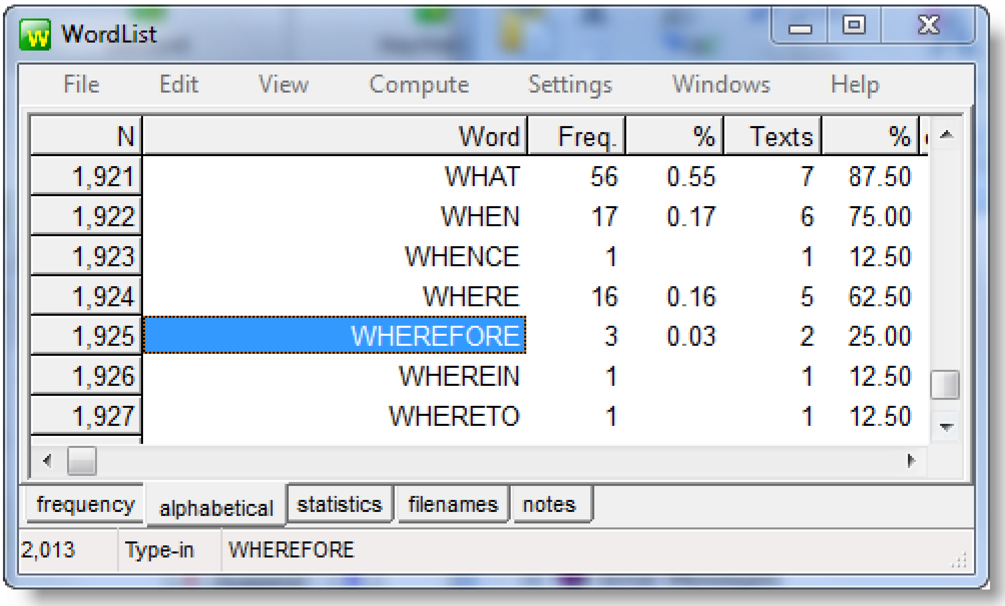
WordList

Section V

5 WordList

5.1 overview

A word list in WordSmith Tools looks something like this:



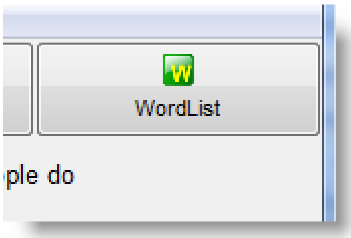
N	Word	Freq.	%	Texts	%
1,921	WHAT	56	0.55	7	87.50
1,922	WHEN	17	0.17	6	75.00
1,923	WHENCE	1		1	12.50
1,924	WHERE	16	0.16	5	62.50
1,925	WHEREFORE	3	0.03	2	25.00
1,926	WHEREIN	1		1	12.50
1,927	WHERETO	1		1	12.50

2,013 Type-in WHEREFORE

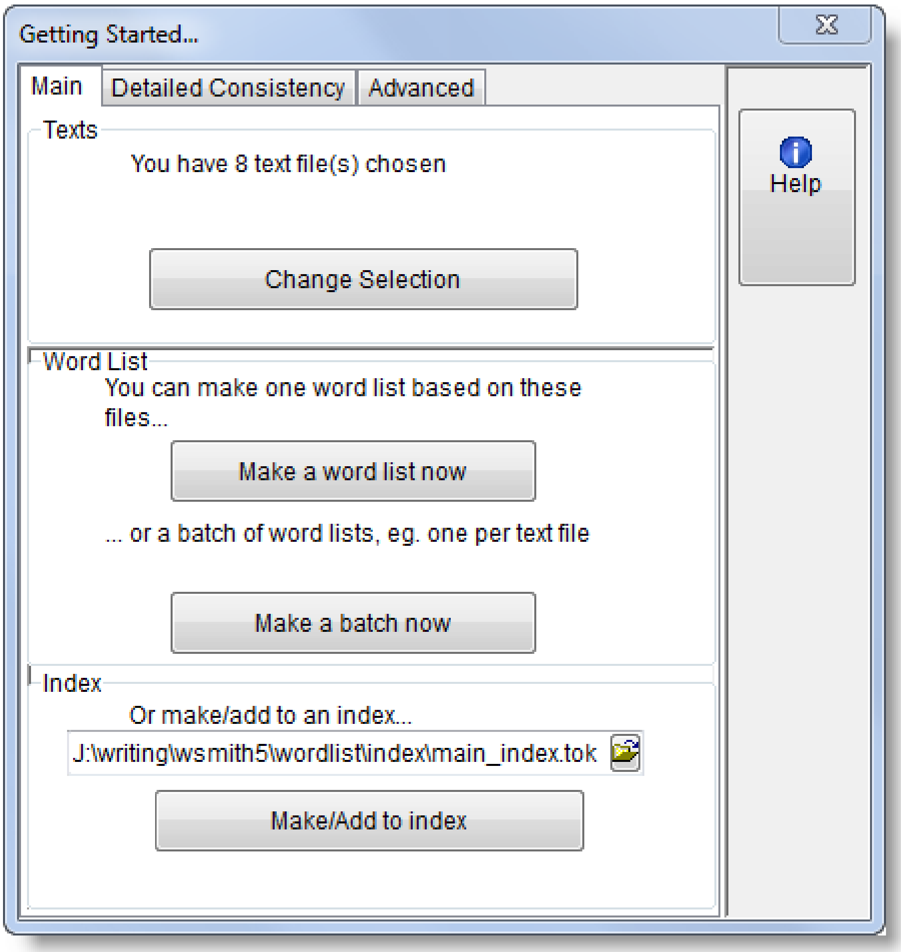
It shows how often each word occurs in the text files, what that is as a percent of the running words in the text, and how many text files each word was found in.

5.2 making a word list

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



When WordList starts up, choose your texts and then you will see something like this.



Getting Started...

Texts

You have 8 text file(s) chosen

Word List

You can make one word list based on these files...

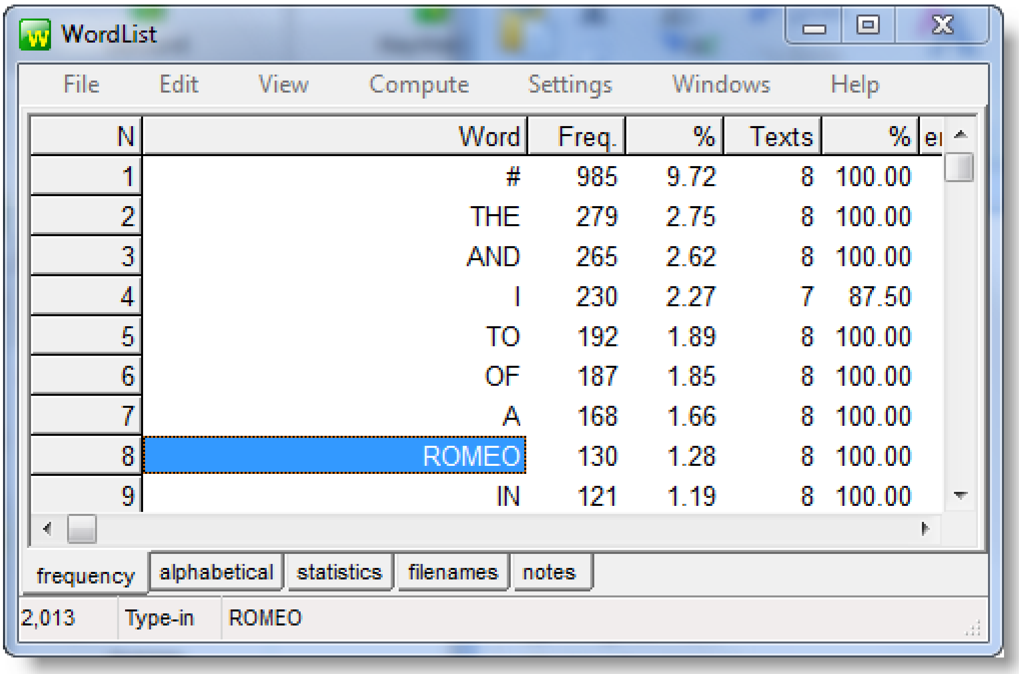
... or a batch of word lists, eg. one per text file

Index

Or make/add to an index...

J:\writing\wsmith5\wordlist\index\main_index.tok

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*.

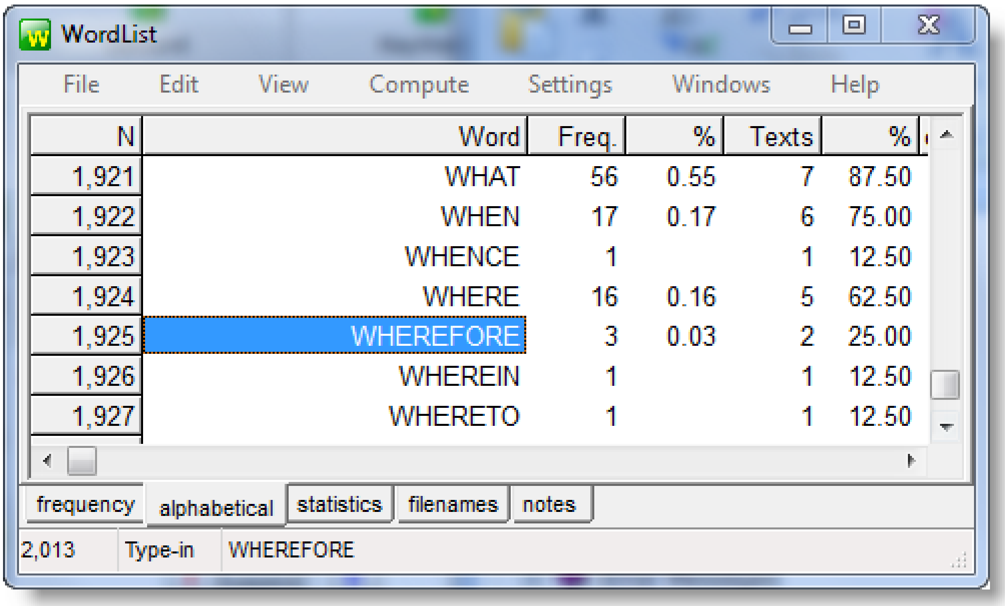


N	Word	Freq.	%	Texts	%
1	#	985	9.72	8	100.00
2	THE	279	2.75	8	100.00
3	AND	265	2.62	8	100.00
4	I	230	2.27	7	87.50
5	TO	192	1.89	8	100.00
6	OF	187	1.85	8	100.00
7	A	168	1.66	8	100.00
8	ROMEO	130	1.28	8	100.00
9	IN	121	1.19	8	100.00

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.



N	Word	Freq.	%	Texts	%
1,921	WHAT	56	0.55	7	87.50
1,922	WHEN	17	0.17	6	75.00
1,923	WHENCE	1		1	12.50
1,924	WHERE	16	0.16	5	62.50
1,925	WHEREFORE	3	0.03	2	25.00
1,926	WHEREIN	1		1	12.50
1,927	WHERETO	1		1	12.50

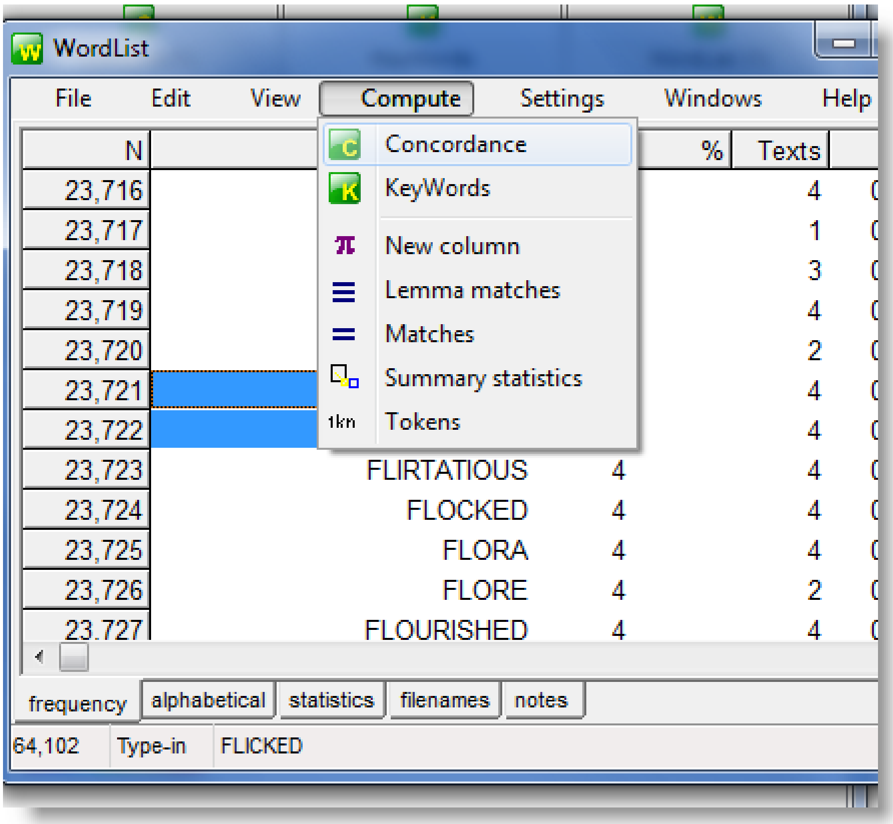
2,013 Type-in WHEREFORE

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)



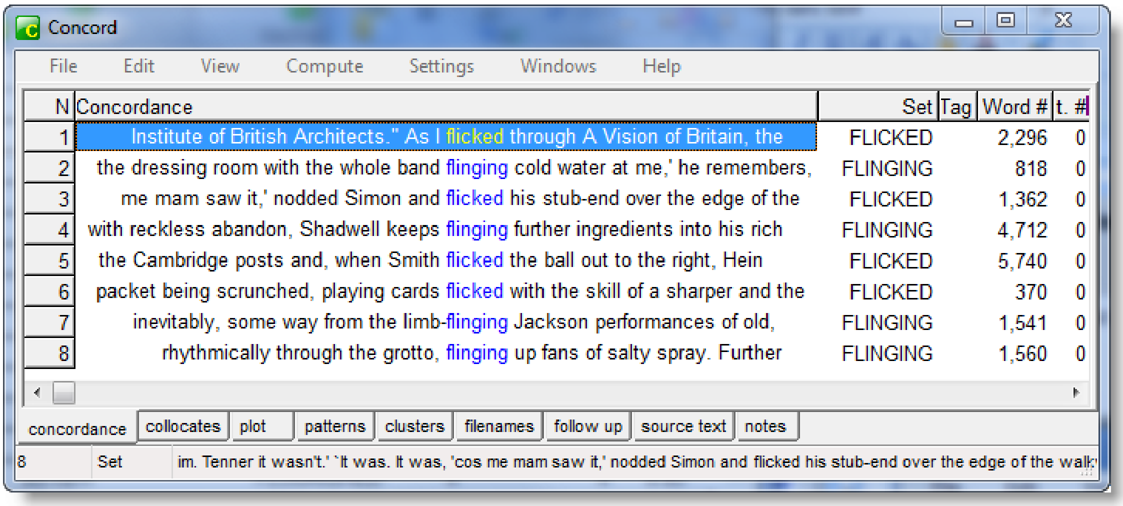
N			%	Texts
23,716				4
23,717				1
23,718				3
23,719				4
23,720				2
23,721				4
23,722				4
23,723	FLIRTATIOUS	4		4
23,724	FLOCKED	4		4
23,725	FLORA	4		4
23,726	FLORE	4		2
23,727	FLOURISHED	4		4

64,102 Type-in FLICKED

e

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):

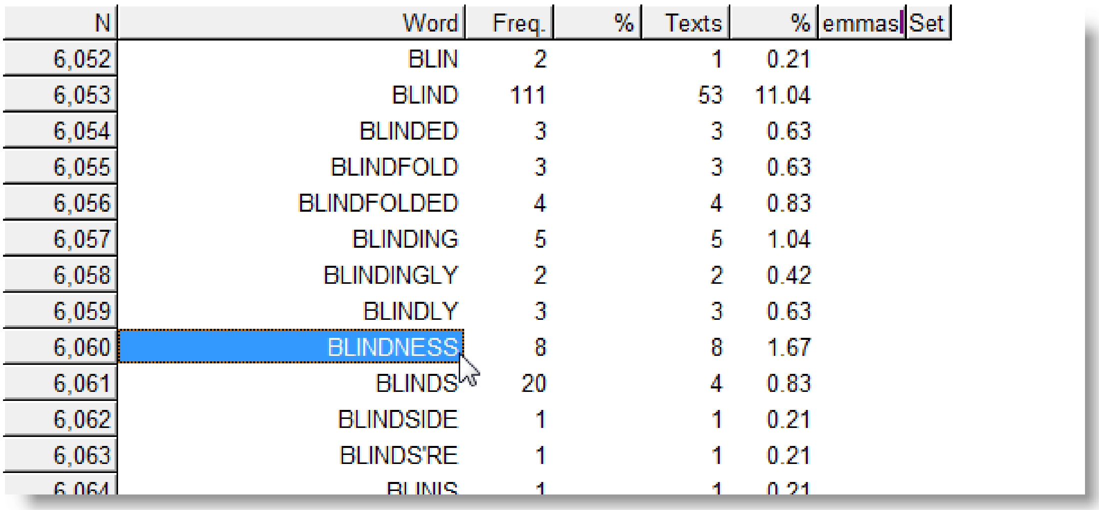


N	Concordance	Set	Tag	Word #	t. #
1	Institute of British Architects." As I flicked through A Vision of Britain, the	FLICKED		2,296	0
2	the dressing room with the whole band flinging cold water at me,' he remembers,	FLINGING		818	0
3	me mam saw it,' nodded Simon and flicked his stub-end over the edge of the	FLICKED		1,362	0
4	with reckless abandon, Shadwell keeps flinging further ingredients into his rich	FLINGING		4,712	0
5	the Cambridge posts and, when Smith flicked the ball out to the right, Hein	FLICKED		5,740	0
6	packet being scrunched, playing cards flicked with the skill of a sharper and the	FLICKED		370	0
7	inevitably, some way from the limb-flinging Jackson performances of old,	FLINGING		1,541	0
8	rhythmically through the grotto, flinging up fans of salty spray. Further	FLINGING		1,560	0

8 Set im. Tenner it wasn't.' `It was. It was, 'cos me mam saw it,' nodded Simon and flicked his stub-end over the edge of the walk

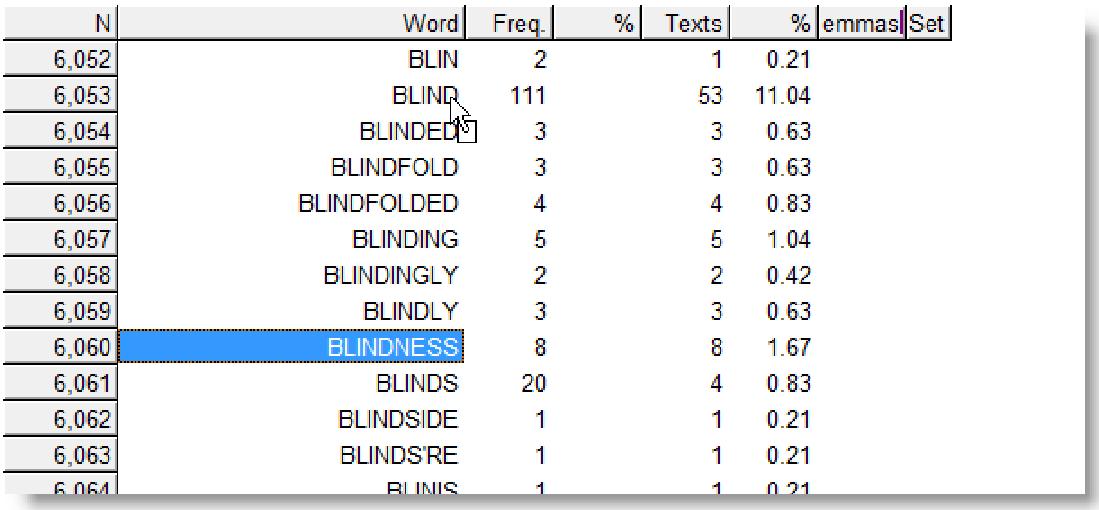
5.4 lemmatising

To lemmatise manually, with a word list on screen,



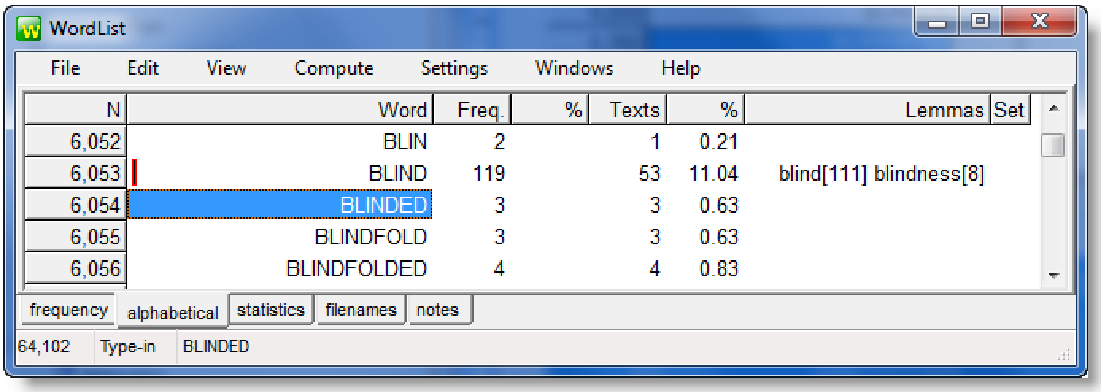
N	Word	Freq.	%	Texts	%	emmas	Set
6,052	BLIN	2		1	0.21		
6,053	BLIND	111		53	11.04		
6,054	BLINDED	3		3	0.63		
6,055	BLINDFOLD	3		3	0.63		
6,056	BLINDFOLDED	4		4	0.83		
6,057	BLINDING	5		5	1.04		
6,058	BLINDINGLY	2		2	0.42		
6,059	BLINDLY	3		3	0.63		
6,060	BLINDNESS	8		8	1.67		
6,061	BLINDS	20		4	0.83		
6,062	BLINDSIDE	1		1	0.21		
6,063	BLINDS'RE	1		1	0.21		
6,064	BLINIS	1		1	0.21		

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



N	Word	Freq.	%	Texts	%	emmas	Set
6,052	BLIN	2		1	0.21		
6,053	BLIND	111		53	11.04		
6,054	BLINDED	3		3	0.63		
6,055	BLINDFOLD	3		3	0.63		
6,056	BLINDFOLDED	4		4	0.83		
6,057	BLINDING	5		5	1.04		
6,058	BLINDINGLY	2		2	0.42		
6,059	BLINDLY	3		3	0.63		
6,060	BLINDNESS	8		8	1.67		
6,061	BLINDS	20		4	0.83		
6,062	BLINDSIDE	1		1	0.21		
6,063	BLINDS'RE	1		1	0.21		
6,064	BLINIS	1		1	0.21		

and drop it:

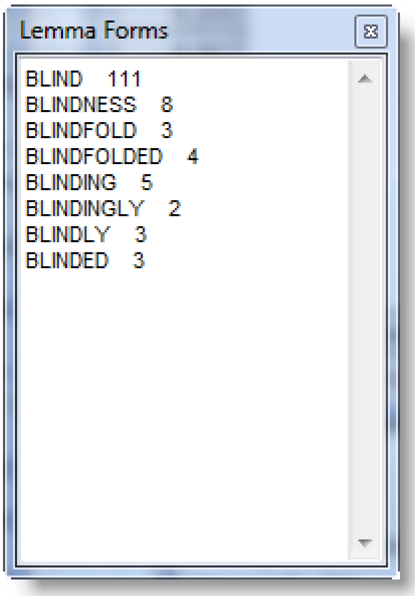


N	Word	Freq.	%	Texts	%	Lemmas	Set
6,052	BLIN	2		1	0.21		
6,053	BLIND	119		53	11.04	blind[111] blindness[8]	
6,054	BLINDED	3		3	0.63		
6,055	BLINDFOLD	3		3	0.63		
6,056	BLINDFOLDED	4		4	0.83		

64,102 Type-in BLINDED

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

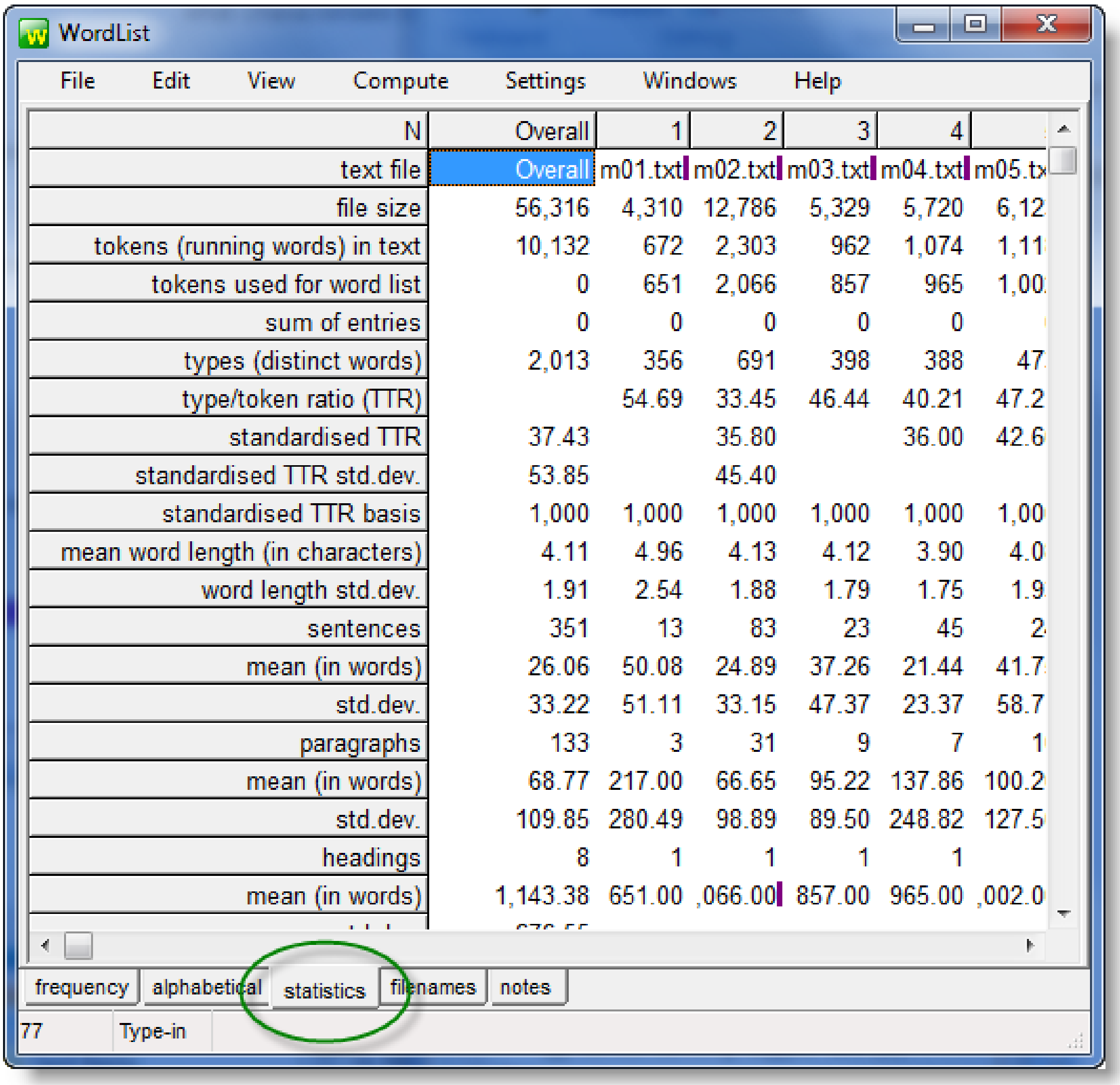
If there are a lot, you can double-click the Lemmas column to see the details:



BLIND 111
BLINDNESS 8
BLINDFOLD 3
BLINDFOLDED 4
BLINDING 5
BLINDINGLY 2
BLINDLY 3
BLINDED 3

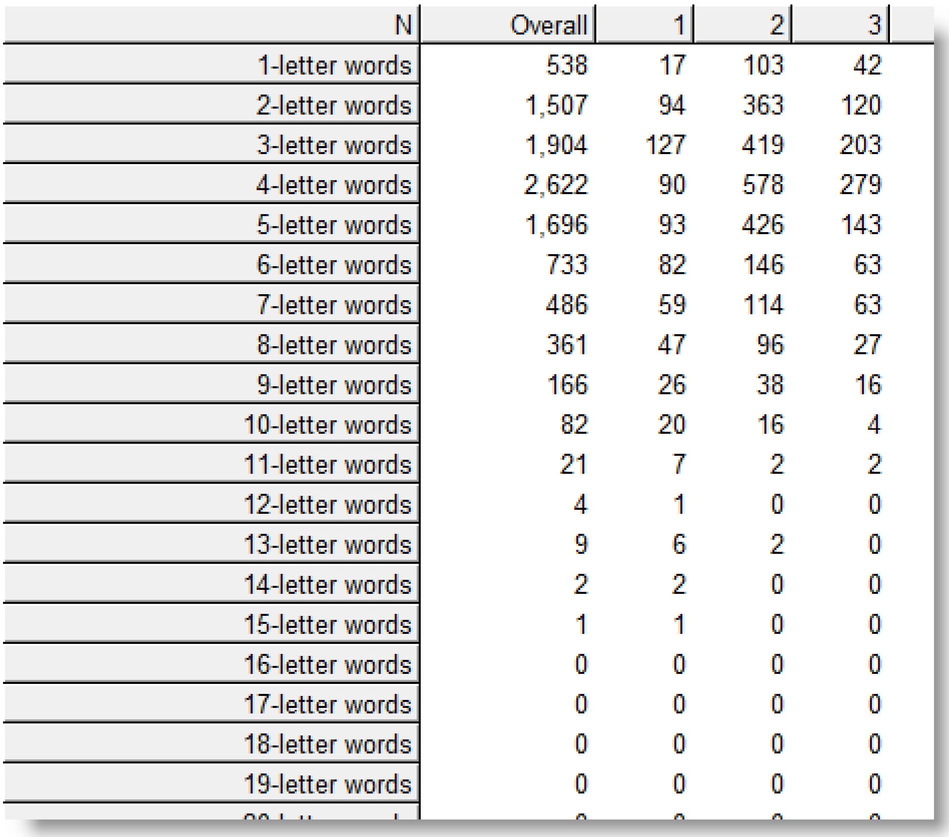
5.5 word list statistics

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



N	Overall	1	2	3	4	
text file	Overall	m01.txt	m02.txt	m03.txt	m04.txt	m05.tx
file size	56,316	4,310	12,786	5,329	5,720	6,12
tokens (running words) in text	10,132	672	2,303	962	1,074	1,11
tokens used for word list	0	651	2,066	857	965	1,00
sum of entries	0	0	0	0	0	
types (distinct words)	2,013	356	691	398	388	47
type/token ratio (TTR)		54.69	33.45	46.44	40.21	47.2
standardised TTR	37.43		35.80		36.00	42.6
standardised TTR std.dev.	53.85		45.40			
standardised TTR basis	1,000	1,000	1,000	1,000	1,000	1,00
mean word length (in characters)	4.11	4.96	4.13	4.12	3.90	4.0
word length std.dev.	1.91	2.54	1.88	1.79	1.75	1.9
sentences	351	13	83	23	45	2
mean (in words)	26.06	50.08	24.89	37.26	21.44	41.7
std.dev.	33.22	51.11	33.15	47.37	23.37	58.7
paragraphs	133	3	31	9	7	1
mean (in words)	68.77	217.00	66.65	95.22	137.86	100.2
std.dev.	109.85	280.49	98.89	89.50	248.82	127.5
headings	8	1	1	1	1	
mean (in words)	1,143.38	651.00	,066.00	857.00	965.00	,002.0

and something like this should appear. Lots of numbers. Further down, the numbers are easier to understand:



N	Overall	1	2	3
1-letter words	538	17	103	42
2-letter words	1,507	94	363	120
3-letter words	1,904	127	419	203
4-letter words	2,622	90	578	279
5-letter words	1,696	93	426	143
6-letter words	733	82	146	63
7-letter words	486	59	114	63
8-letter words	361	47	96	27
9-letter words	166	26	38	16
10-letter words	82	20	16	4
11-letter words	21	7	2	2
12-letter words	4	1	0	0
13-letter words	9	6	2	0
14-letter words	2	2	0	0
15-letter words	1	1	0	0
16-letter words	0	0	0	0
17-letter words	0	0	0	0
18-letter words	0	0	0	0
19-letter words	0	0	0	0

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

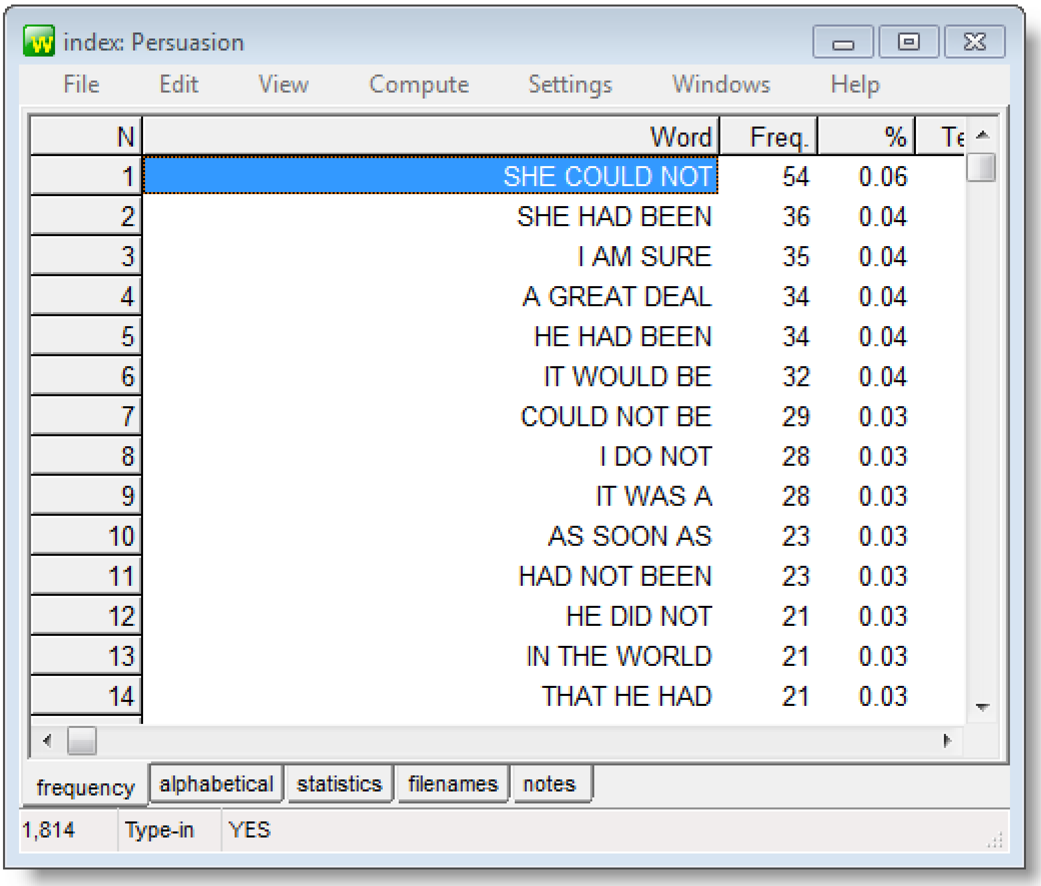
```
OF THE
IN THE END
ONCE UPON A TIME
```

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.



N	Word	Freq.	%	Te
1	SHE COULD NOT	54	0.06	
2	SHE HAD BEEN	36	0.04	
3	I AM SURE	35	0.04	
4	A GREAT DEAL	34	0.04	
5	HE HAD BEEN	34	0.04	
6	IT WOULD BE	32	0.04	
7	COULD NOT BE	29	0.03	
8	I DO NOT	28	0.03	
9	IT WAS A	28	0.03	
10	AS SOON AS	23	0.03	
11	HAD NOT BEEN	23	0.03	
12	HE DID NOT	21	0.03	
13	IN THE WORLD	21	0.03	
14	THAT HE HAD	21	0.03	

1,814 Type-in YES

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

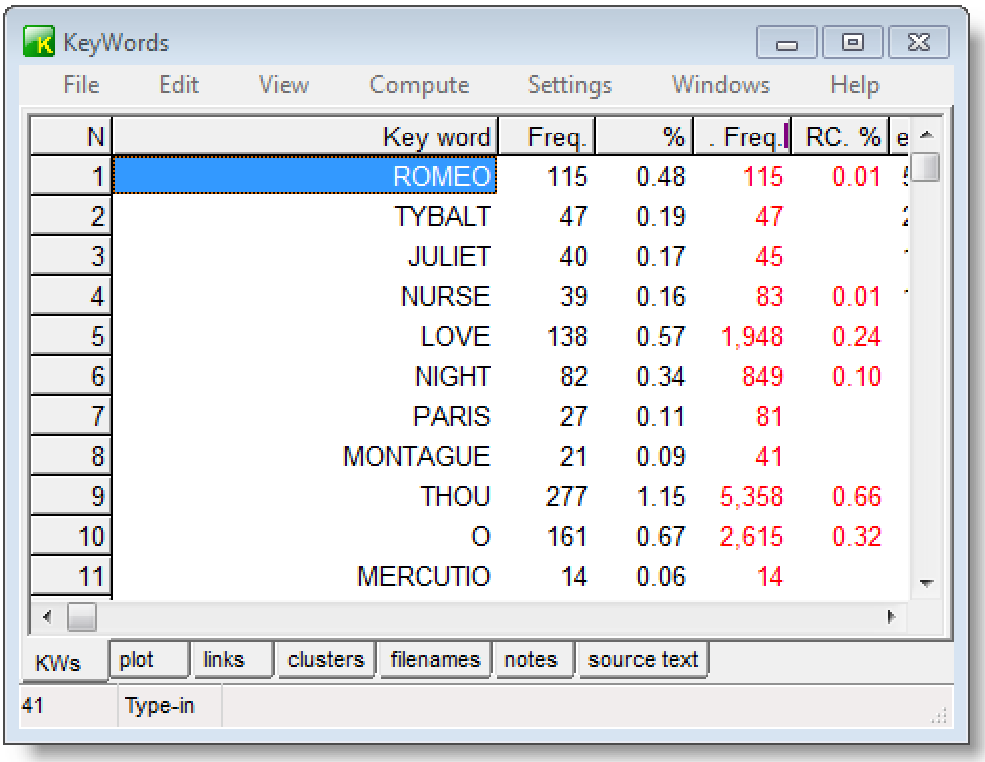
KeyWords

Section VI

6 KeyWords

6.1 overview

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



N	Key word	Freq.	%	. Freq.	RC. %	e
1	ROMEO	115	0.48	115	0.01	5
2	TYBALT	47	0.19	47		2
3	JULIET	40	0.17	45		1
4	NURSE	39	0.16	83	0.01	1
5	LOVE	138	0.57	1,948	0.24	
6	NIGHT	82	0.34	849	0.10	
7	PARIS	27	0.11	81		
8	MONTAGUE	21	0.09	41		
9	THOU	277	1.15	5,358	0.66	
10	O	161	0.67	2,615	0.32	
11	MERCUTIO	14	0.06	14		

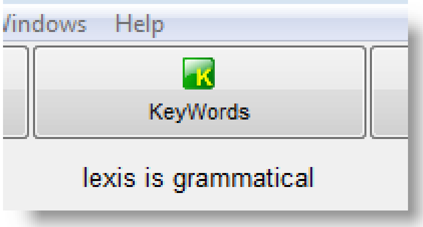
The key words are words which occur unusually frequently in comparison with some kind of reference corpus.

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.

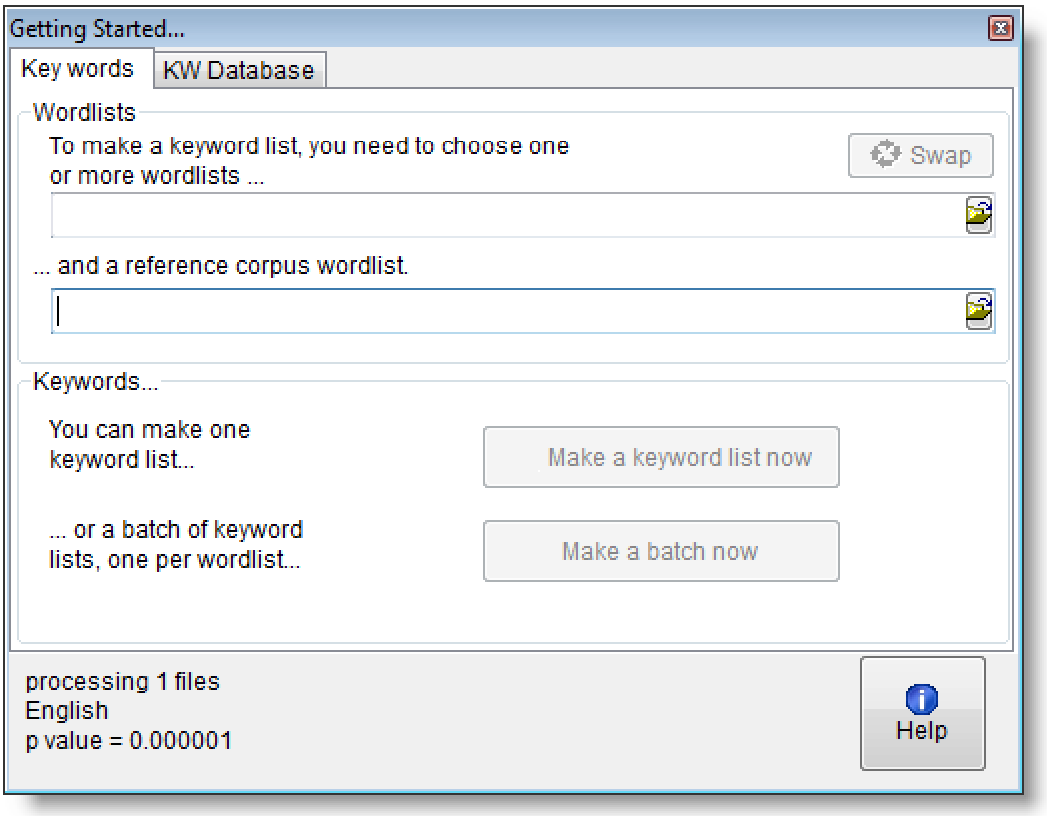
In the list above, based on the play Romeo and Juliet in comparison with all the Shakespeare plays, we see lots of names of the main characters, some pronouns like **`thou`**, plus theme words like **`love`** and **`night`**.

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.



To make a keyword list, you need to choose one or more wordlists ...

... and a reference corpus wordlist.

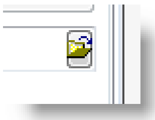
You can make one keyword list...

... or a batch of keyword lists, one per wordlist...

processing 1 files
English
p value = 0.000001

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

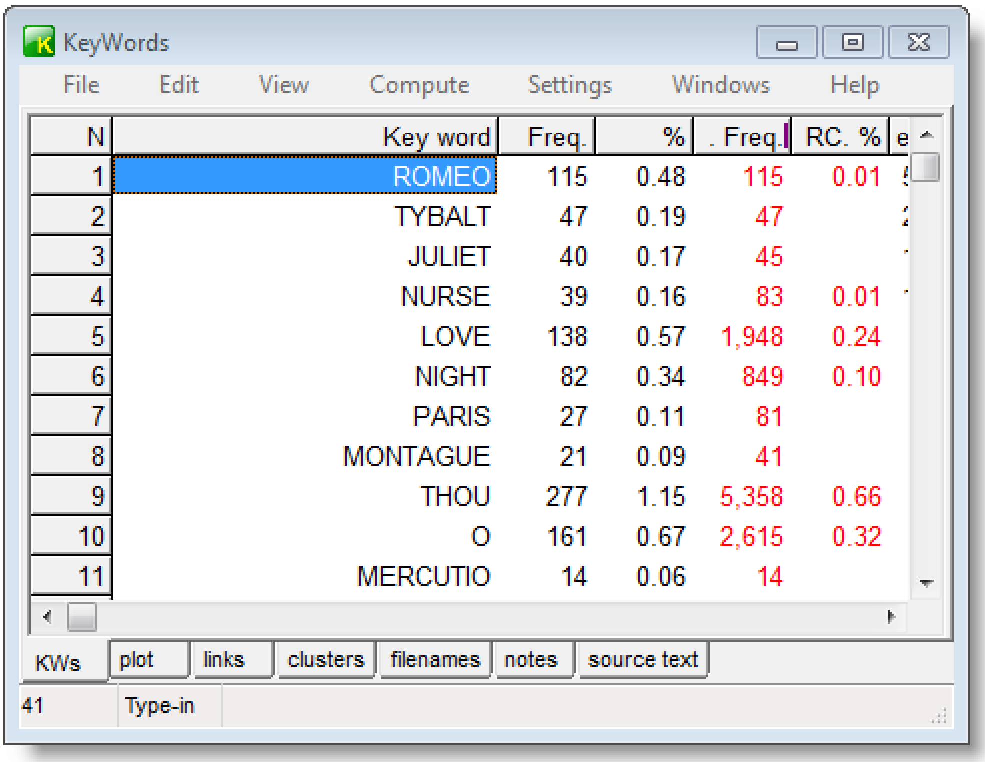
You can choose the word list files by pressing this button:



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.)

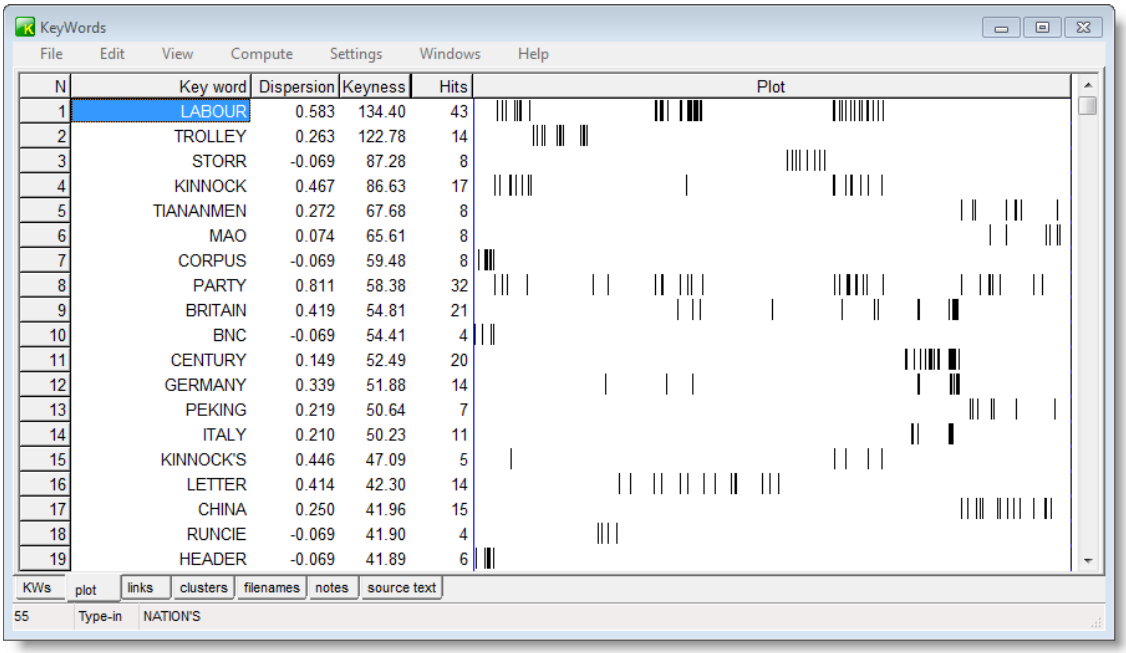
Then you will see something like this:



N	Key word	Freq.	%	. Freq.	RC. %	e
1	ROMEO	115	0.48	115	0.01	5
2	TYBALT	47	0.19	47		2
3	JULIET	40	0.17	45		1
4	NURSE	39	0.16	83	0.01	1
5	LOVE	138	0.57	1,948	0.24	
6	NIGHT	82	0.34	849	0.10	
7	PARIS	27	0.11	81		
8	MONTAGUE	21	0.09	41		
9	THOU	277	1.15	5,358	0.66	
10	O	161	0.67	2,615	0.32	
11	MERCUTIO	14	0.06	14		

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.



N	Key word	Dispersion	Keyness	Hits	Plot
1	LABOUR	0.583	134.40	43	
2	TROLLEY	0.263	122.78	14	
3	STORR	-0.069	87.28	8	
4	KINNOCK	0.467	86.63	17	
5	TIANANMEN	0.272	67.68	8	
6	MAO	0.074	65.61	8	
7	CORPUS	-0.069	59.48	8	
8	PARTY	0.811	58.38	32	
9	BRITAIN	0.419	54.81	21	
10	BNC	-0.069	54.41	4	
11	CENTURY	0.149	52.49	20	
12	GERMANY	0.339	51.88	14	
13	PEKING	0.219	50.64	7	
14	ITALY	0.210	50.23	11	
15	KINNOCK'S	0.446	47.09	5	
16	LETTER	0.414	42.30	14	
17	CHINA	0.250	41.96	15	
18	RUNCIE	-0.069	41.90	4	
19	HEADER	-0.069	41.89	6	

55 Type-in NATION'S

You see:

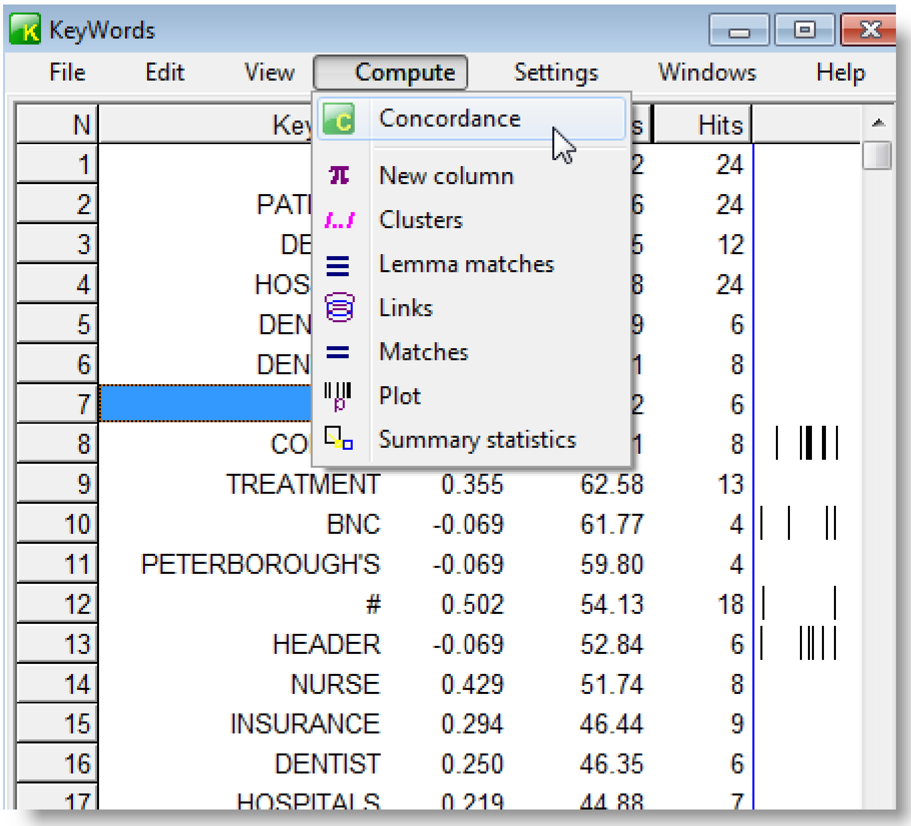
- each key word (KW) (these obviously have to do with international relations)
- a measure of its dispersion and its keyness
- how many times each KW came in the text (hits).
- a map showing where each word came.

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.

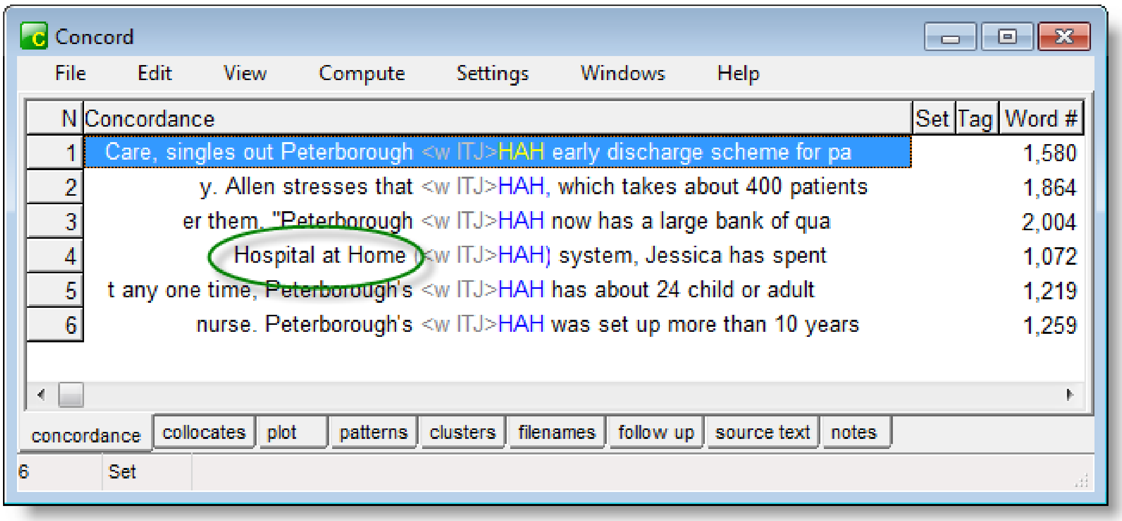
Select a word (or more)



N	Key		s	Hits
1			2	24
2	PATI		6	24
3	DE		5	12
4	HOS		8	24
5	DEN		9	6
6	DEN		1	8
7			2	6
8	CO		1	8
9	TREATMENT	0.355	62.58	13
10	BNC	-0.069	61.77	4
11	PETERBOROUGH'S	-0.069	59.80	4
12	#	0.502	54.13	18
13	HEADER	-0.069	52.84	6
14	NURSE	0.429	51.74	8
15	INSURANCE	0.294	46.44	9
16	DENTIST	0.250	46.35	6
17	HOSPITALS	0.219	44.88	7

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):



N	Concordance	Set	Tag	Word #
1	Care, singles out Peterborough <w ITJ>HAH early discharge scheme for pa			1,580
2	y. Allen stresses that <w ITJ>HAH, which takes about 400 patients			1,864
3	er them. "Peterborough <w ITJ>HAH now has a large bank of qua			2,004
4	Hospital at Home <w ITJ>HAH) system, Jessica has spent			1,072
5	t any one time, Peterborough's <w ITJ>HAH has about 24 child or adult			1,219
6	nurse. Peterborough's <w ITJ>HAH was set up more than 10 years			1,259

6 Set

It seems that HAH is a system for health care.

Index