

# Step-by-step guide to WordSmith



Version 9.0

© 2024 Mike Scott

<%PUBLISHER%>

<%PUBLISHER-CITY%>

# ***Step-by-step guide to WordSmith***

version 9.0

---

*by Mike Scott*

Compiled: July, 2024

# Step-by-step guide to WordSmith

© 2024 Mike Scott

All rights reserved. But most parts of this work may be reproduced in any form or by any means - graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems - usually without the written permission of the publisher.

See [http://www.lexically.net/publications/copyright\\_permission\\_for\\_screenshots.htm](http://www.lexically.net/publications/copyright_permission_for_screenshots.htm)

Products that are referred to in this document may be either trademarks and/or registered trademarks of the respective owners. The publisher and the author make no claim to these trademarks.

While every precaution has been taken in the preparation of this document, the publisher and the author assume no responsibility for errors or omissions, or for damages resulting from the use of information contained in this document or from the use of programs and source code that may accompany it. In no event shall the publisher and the author be liable for any loss of profit or any other commercial damage caused or alleged to have been caused directly or indirectly by this document.

Produced: July 2024

## Publisher

<%Publisher%>

## Special thanks to:

*All the people who contributed to this document by testing WordSmith Tools in its various incarnations. Especially those who reported problems and sent me suggestions.*

# Table of Contents

Foreword	0
<b>Part I Introduction</b>	<b>1</b>
<b>Part II Choosing your texts</b>	<b>4</b>
<b>Part III Select the right language</b>	<b>9</b>
<b>Part IV Concordancing</b>	<b>12</b>
1 overview .....	13
2 making a concordance .....	13
3 seeing the source text .....	17
4 collocates and mutual information .....	19
5 concordancing tagged text (1) .....	20
6 concordancing tagged text (2) .....	25
<b>Part V WordList</b>	<b>28</b>
1 overview .....	29
2 making a word list .....	30
3 concordancing selected words .....	32
4 lemmatising .....	34
5 word list statistics .....	35
6 multi-word units .....	35
using an index .....	35
making a multi-word wordlist .....	36
<b>Part VI KeyWords</b>	<b>37</b>
1 overview .....	38
2 making a key word list .....	39
3 key words plot .....	43
4 concordancing selected key words .....	43
<b>Index</b>	<b>45</b>

*Step-by-step guide to WordSmith*

# *Introduction*

**Section**

---



**/**

# 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 three buttons for the main Tools, and a series of tabs below for adjusting settings.

Concord makes concordances, KeyWords finds the key words in texts, and WordList generates lists of the words in a text or a set of texts. To get started, press one of those three buttons.




*Step-by-step guide to WordSmith*

# *Choosing your texts*

**Section**

---

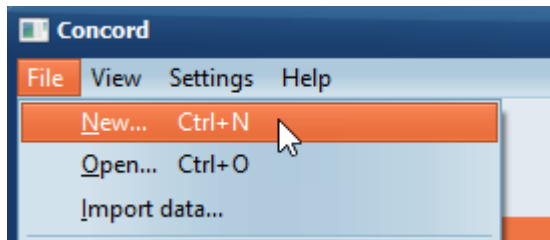


//

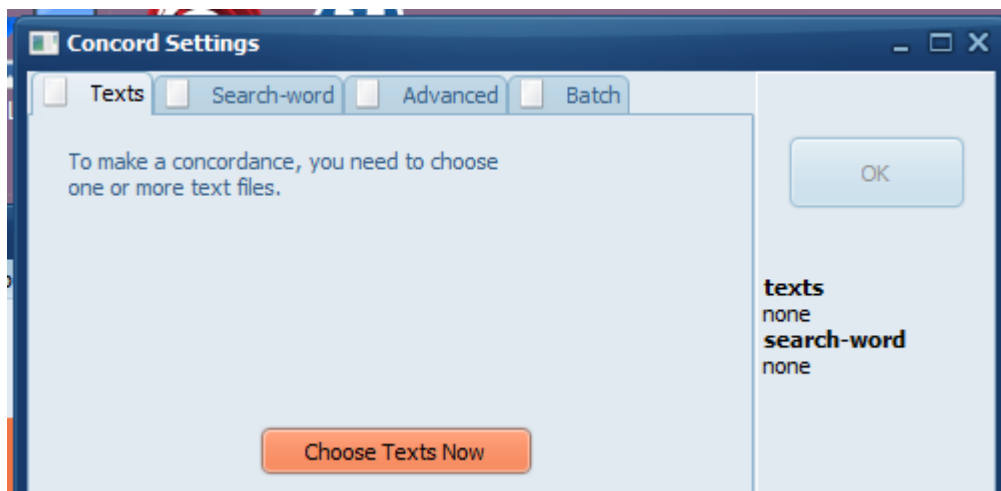


## 2 Choosing your texts

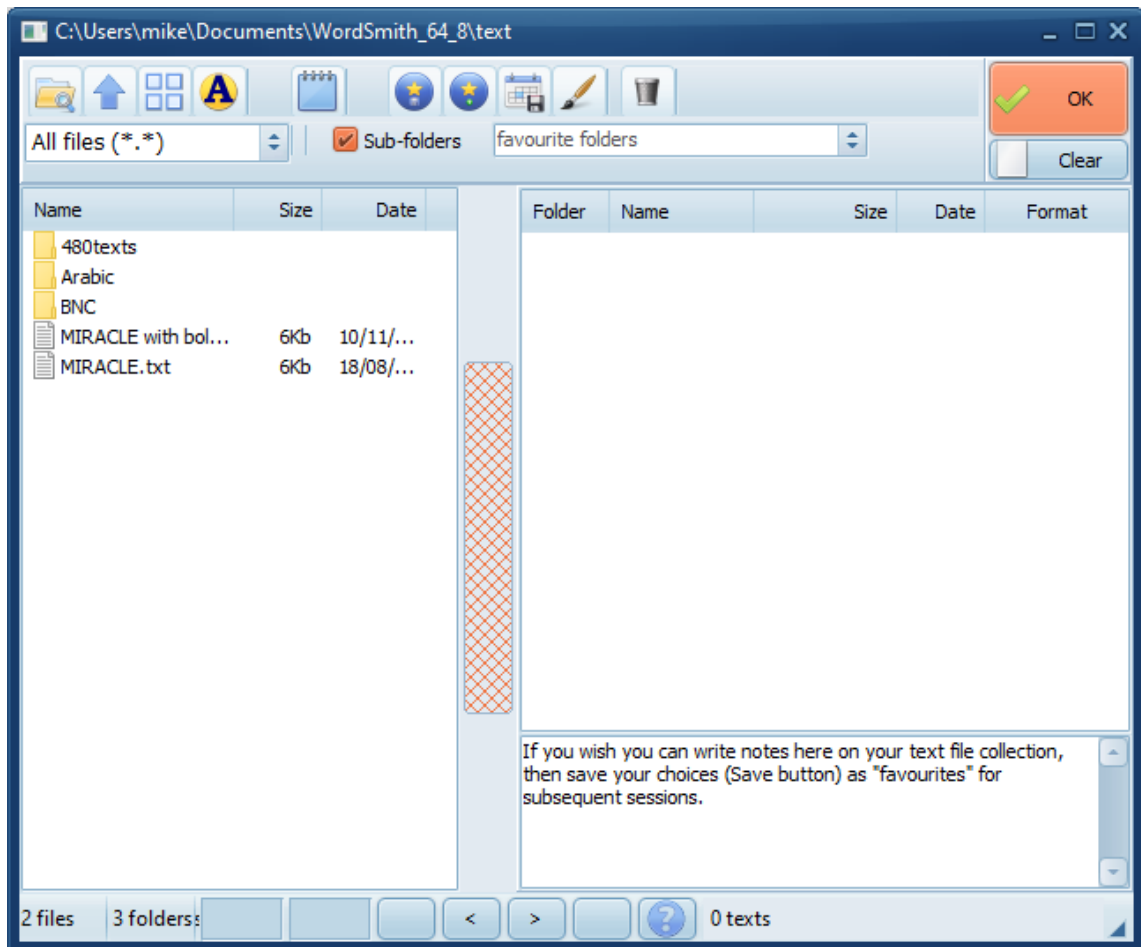
To choose text files, click the *File* | *New* menu option in the Tool you want to use:




and you get to a window with a button asking for you to choose text(s).

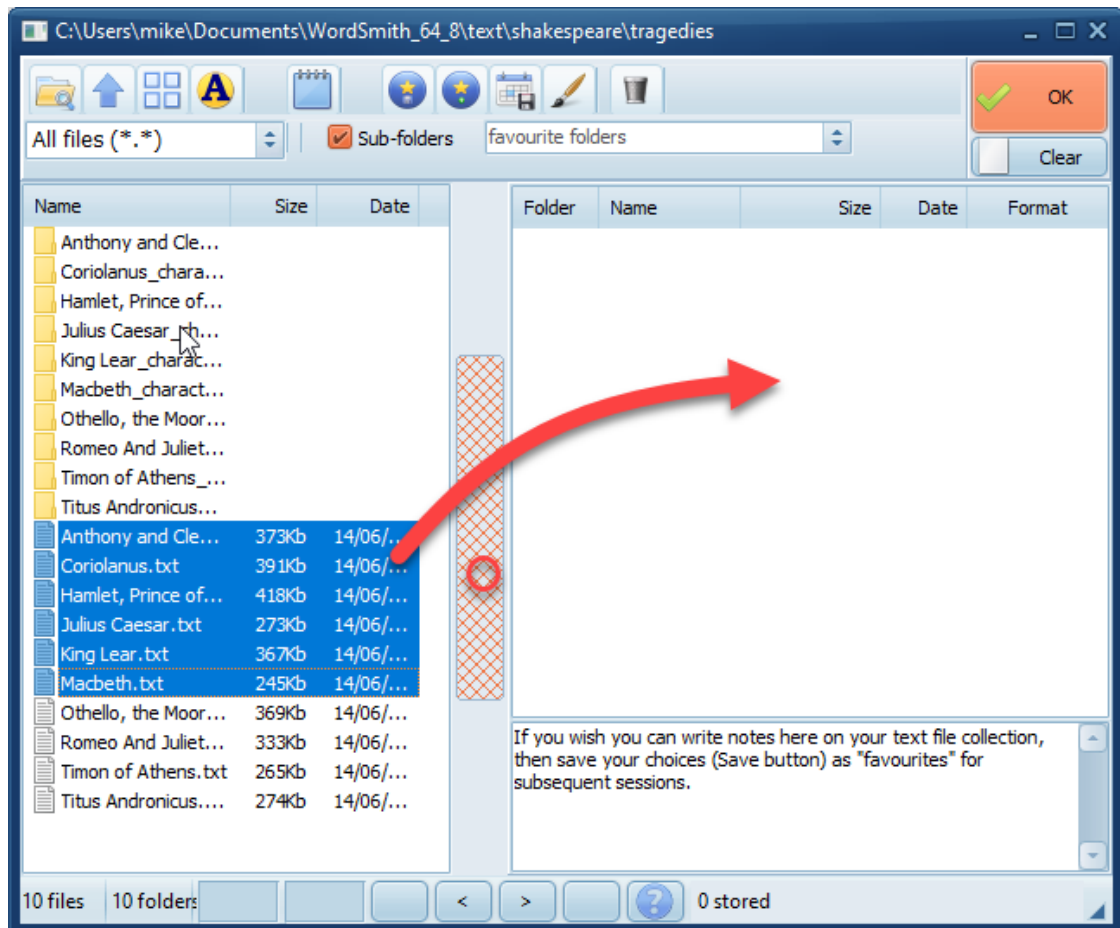


When you click *Choose Texts*, you will see something like this:

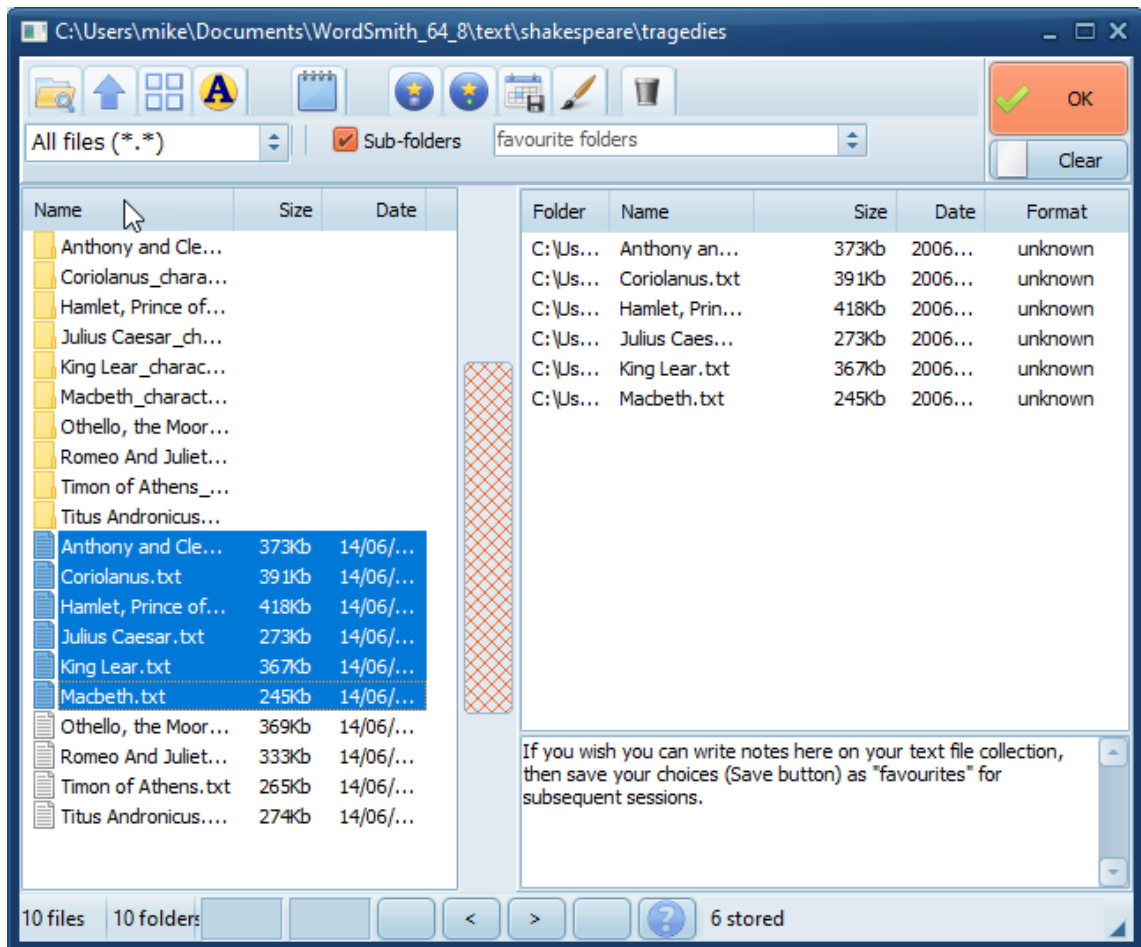


At the left is a fairly standard text file explorer, at the right an area for Files selected.

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



Click the red checked button, 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 text files have been stored.

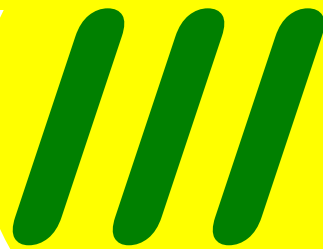
Press the OK button.

*Step-by-step guide to WordSmith*

# *Select the right language*

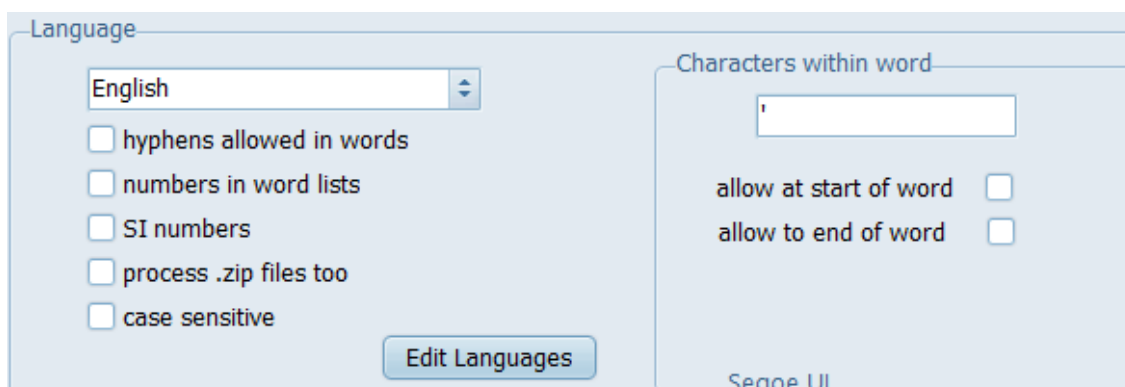
**Section**

---

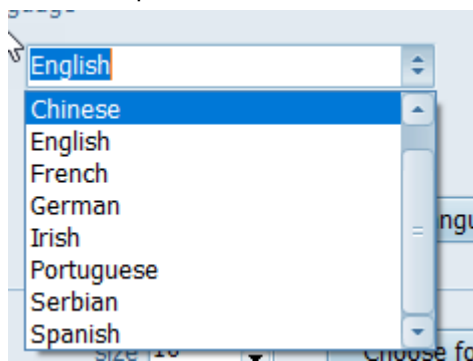


### 3 Select the right language

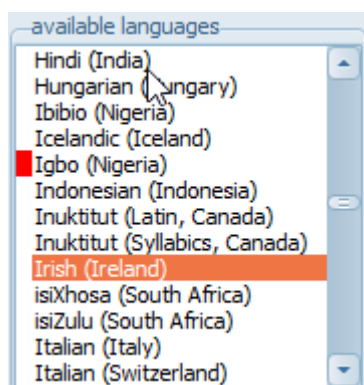
Most examples in this guide deal with texts in English. If you want to handle texts in Chinese or some other language, you need to choose the language in the main Controller.



If in the drop-down list



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.

chosen languages

Name	Status	Font	S...	Word...
English	#1	Sego...	10	'
French	*	Sego...	12	
German	*	Sego...	10	
Igbo	*	Sego...	10	'
Irish	*	Sego...	10	
Portuguese	*	Sego...	12	
Serbian	*	Sego...	10	'
Spanish	*	Sego...	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.

Language Chooser

File Edit

☐ main language ☒ available language

Igbo

Font

Segoe UI

10

Segoe UI

Treatment

extra characters

'

allow at start of word ☐

allow to right end of word ☐

hyphens allowed in words ☐

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.

Language Chooser

File Edit

Load custom data

Save

Exit Alt+X

*Step-by-step guide to WordSmith*

# *Concordancing*

**Section**

---

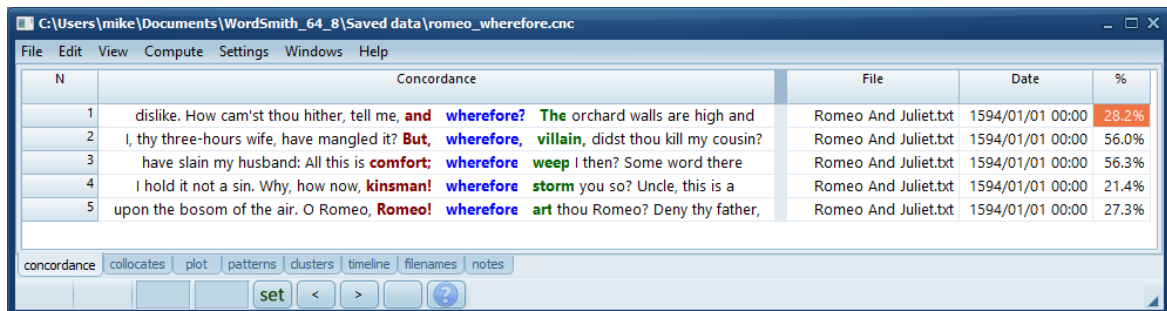
**IV**



## 4 Concordancing

### 4.1 overview

A concordance looks something like this:

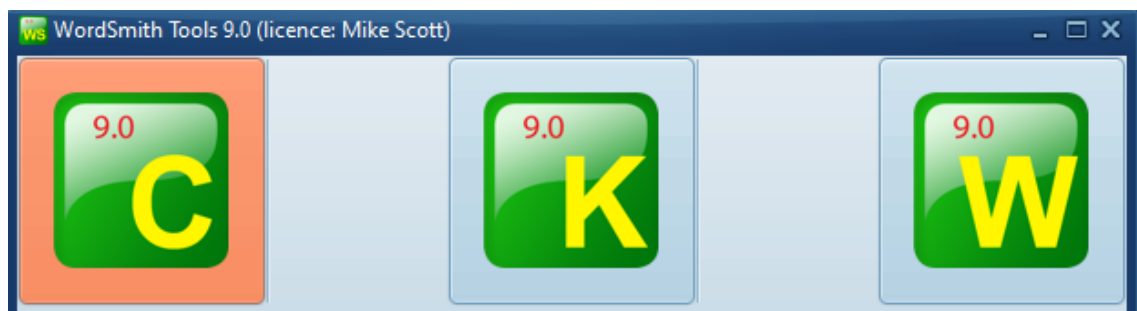


N	Concordance	File	Date	%
1	dislike. How cam'st thou hither, tell me, <b>and wherefore?</b> The orchard walls are high and	Romeo And Juliet.txt	1594/01/01 00:00	28.2%
2	I, thy three-hours wife, have mangled it? <b>But, wherefore, villain,</b> didst thou kill my cousin?	Romeo And Juliet.txt	1594/01/01 00:00	56.0%
3	have slain my husband: All this is <b>comfort; wherefore weep</b> I then? Some word there	Romeo And Juliet.txt	1594/01/01 00:00	56.3%
4	I hold it not a sin. Why, how now, <b>kinsman! wherefore storm</b> you so? Uncle, this is a	Romeo And Juliet.txt	1594/01/01 00:00	21.4%
5	upon the bosom of the air. O Romeo, <b>Romeo! wherefore art</b> thou Romeo? Deny thy father,	Romeo And Juliet.txt	1594/01/01 00:00	27.3%

It's a concordance of all the occurrences of **wherefore** in Romeo and Juliet. There are only 5 occurrences of **wherefore**.

### 4.2 making a concordance

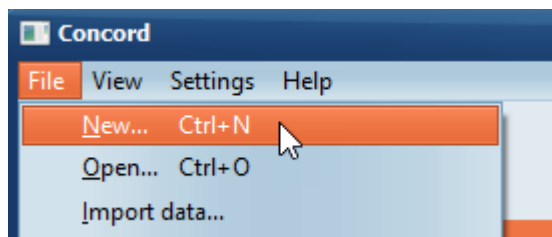
When you press the Concord button in the main Controller,



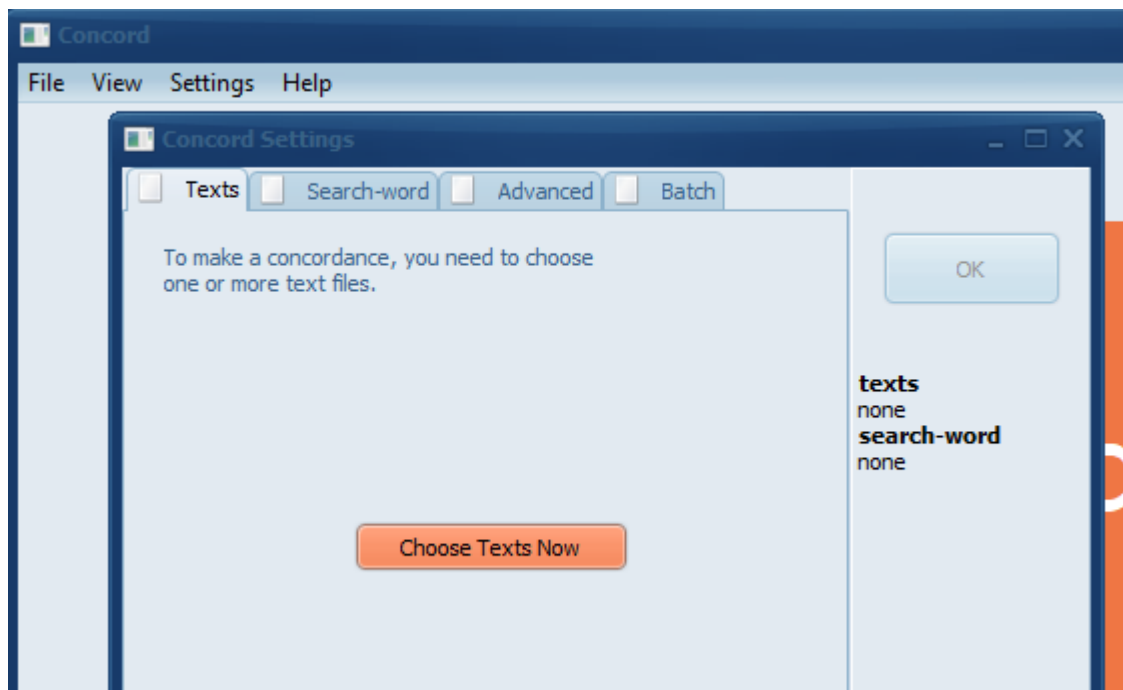
a new Concord Tool opens up and will be visible in the Windows Taskbar.



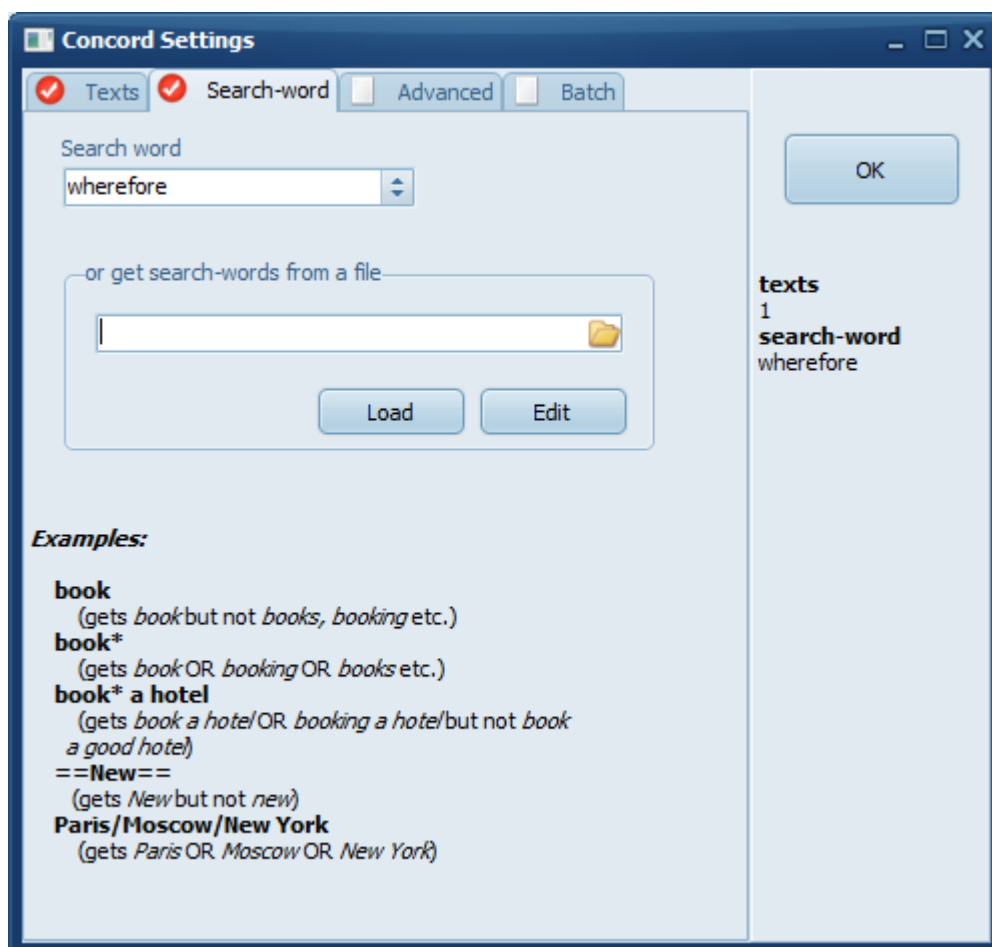
Now in Concord itself, choose *File* | *New*.



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.

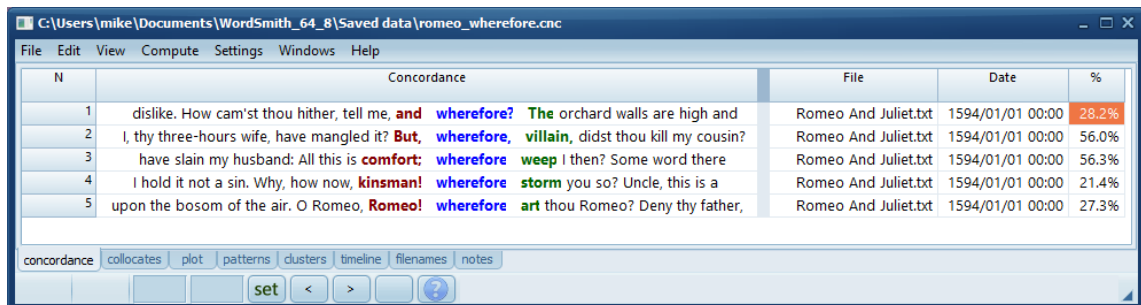
N	Concordance	File	Date	%
1	dislike. How cam'st thou hither, tell me, <b>and wherefore?</b> The orchard walls are high and	Romeo And Juliet.txt	1594/01/01 00:00	28.2%
2	I, thy three-hours wife, have mangled it? <b>But, wherefore, villain,</b> didst thou kill my cousin?	Romeo And Juliet.txt	1594/01/01 00:00	56.0%
3	have slain my husband: All this is <b>comfort; wherefore weep</b> I then? Some word there	Romeo And Juliet.txt	1594/01/01 00:00	56.3%
4	I hold it not a sin. Why, how now, <b>kinsman! wherefore storm</b> you so? Uncle, this is a	Romeo And Juliet.txt	1594/01/01 00:00	21.4%
5	upon the bosom of the air. O Romeo, <b>Romeo! wherefore art</b> thou Romeo? Deny thy father,	Romeo And Juliet.txt	1594/01/01 00:00	27.3%

concordance collocates plot patterns clusters timeline filenames notes

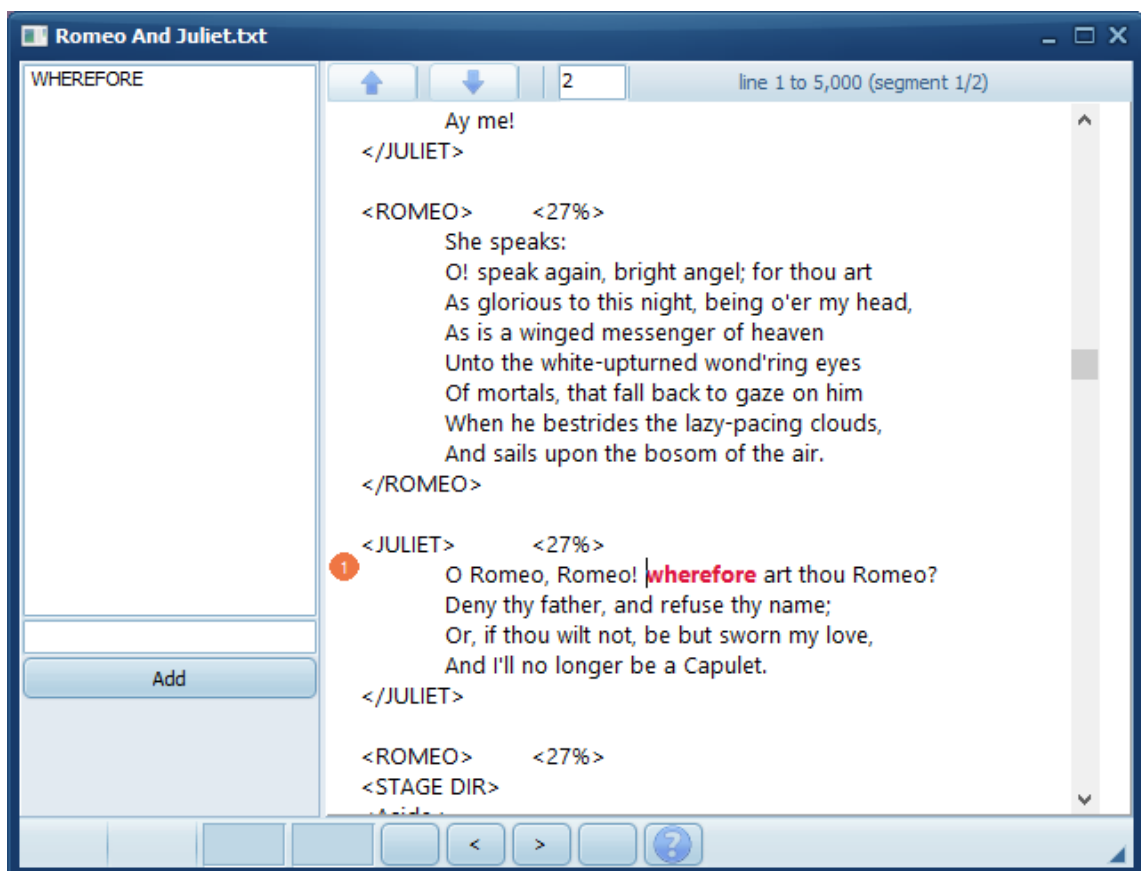
The concordance lists all the examples of "wherefore" which had a word-separator such as punctuation, space etc. before and after it.

### 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	File	Date	%
1	dislike. How cam'st thou hither, tell me, <b>and</b> <b>wherefore?</b> <b>The</b> orchard walls are high and	Romeo And Juliet.txt	1594/01/01 00:00	28.2%
2	I, thy three-hours wife, have mangled it? <b>But</b> , <b>wherefore</b> , <b>villain</b> , didst thou kill my cousin?	Romeo And Juliet.txt	1594/01/01 00:00	56.0%
3	have slain my husband: All this is <b>comfort</b> ; <b>wherefore</b> <b>weep</b> I then? Some word there	Romeo And Juliet.txt	1594/01/01 00:00	56.3%
4	I hold it not a sin. Why, how now, <b>kinsman!</b> <b>wherefore</b> <b>storm</b> you so? Uncle, this is a	Romeo And Juliet.txt	1594/01/01 00:00	21.4%
5	upon the bosom of the air. O Romeo, <b>Romeo!</b> <b>wherefore</b> <b>art</b> thou Romeo? Deny thy father,	Romeo And Juliet.txt	1594/01/01 00:00	27.3%



WHEREFORE

Ay me!

</JULIET>

<ROMEO> <27%>

She speaks:

O! speak again, bright angel; for thou art  
As glorious to this night, being o'er my head,  
As is a winged messenger 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.

</ROMEO>

<JULIET> <27%>

1 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%>

<STAGE DIR>

or press F8 and the lines grow fatter:

N	Concordance
1	yet I know the sound: Art thou not Romeo, and a Montague? Neither, fair maid, if either thee dislike. How cam'st thou hither, tell me, <b>and</b> <b>wherefore?</b> <b>The</b> orchard walls are high and hard to climb, And the place death, considering who thou art, If any of my kinsmen find thee
2	speak ill of him that is my husband? Ah! poor my lord, what tongue shall smooth thy name, When I, thy three-hours wife, have mangled it? <b>But</b> , <b>wherefore</b> , <b>villain</b> , didst thou kill my cousin? That villain cousin would have kill'd my husband: Back, foolish tears, back to your native spring;
3	offer up to joy. My husband lives, that Tybalt would have slain; And Tybalt's dead, that would have slain my husband: All this is <b>comfort</b> ; <b>wherefore</b> <b>weep</b> 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
4	To flee and scorn at our solemnity? Now, by the stock and honour of my kin, To strike him dead I hold it not a sin. Why, how now, <b>kinsman!</b> <b>wherefore</b> <b>storm</b> you so? Uncle, this is a Montague, our foe; A villain that is hither come in spite, To scorn at our solemnity this night.
5	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, <b>Romeo!</b> <b>wherefore</b> <b>art</b> 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.

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

N	Concordance
1	dislike. How cam'st thou hither, tell me, <b>and</b> <b>wherefore?</b> <b>The</b> orchard walls are high and
2	I, thy three-hours wife, have mangled it? <b>But</b> , <b>wherefore</b> , <b>villain</b> , didst thou kill my cousin?
3	tears, back to your native spring; Your tributary drops belong to woe, Which you, mistaking, offer up to joy. My husband lives, that Tybalt would have slain; And Tybalt's dead, that would have slain my husband: All this is <b>comfort</b> ; <b>wherefore</b> <b>weep</b> 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 dead, and Romeo banished!' That
4	I hold it not a sin. Why, how now, <b>kinsman!</b> <b>wherefore</b> <b>storm</b> you so? Uncle, this is a
5	upon the bosom of the air. O Romeo, <b>Romeo!</b> <b>wherefore</b> <b>art</b> thou Romeo? Deny thy father,

and pull it down. You can also pull any column wider or narrower.

## 4.4 collocates and mutual information

Here are the collocates of **KISS** computed using Shakespeare plays, ordered by frequency.

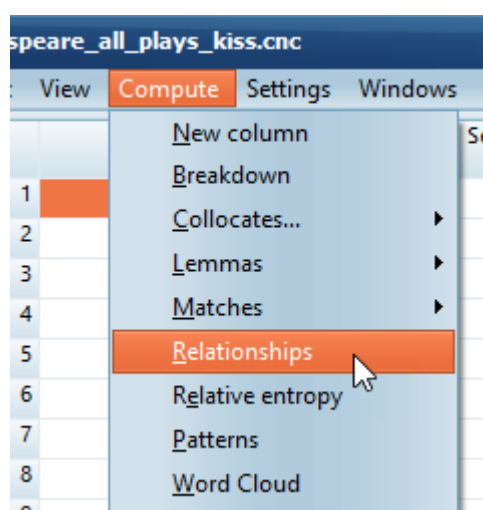
The screenshot shows a software window titled 'shakespeare\_all\_plays\_kiss.cnc'. It displays a table of collocates for the word 'KISS'. The table has columns for 'N', 'Word', 'Texts', 'Total', 'Total Left', 'Total Right', and positions L5, L4, L3, L2, L1, Centre, R1, R2, R3, R4, R5. The data is as follows:

N	Word	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	KISS	33	190	3	3				3		184		3			
2	AND	22	48	24	24	2	3	7	3	9		9	2	10	2	
3	THE	22	30	8	22	5	1	1	1			16	2	4		
4	HAND	16	23	1	22			1					16	5		
5	TO	17	22	18	4	1	1	2		14		2		1		
6	ME	12	17	7	10				3	4		9			1	
7	YOU	9	17	9	8	3	1	2	1	2		3	2		1	
8	THY	12	16	2	14			2				12	1		1	
9	WITH	13	16	10	6	1	1	4	4				2	2	2	
10	YOUR	12	16	2	14			2				12		1		
11	MY	12	15	5	10	4		1				5	2	1		
12	OF	13	14	4	10	1		2	1			3		4		
13	THAT	12	13	8	5	1			2	5		3		2		
14	NOT	10	12	6	6	1		1	2	2				1	3	

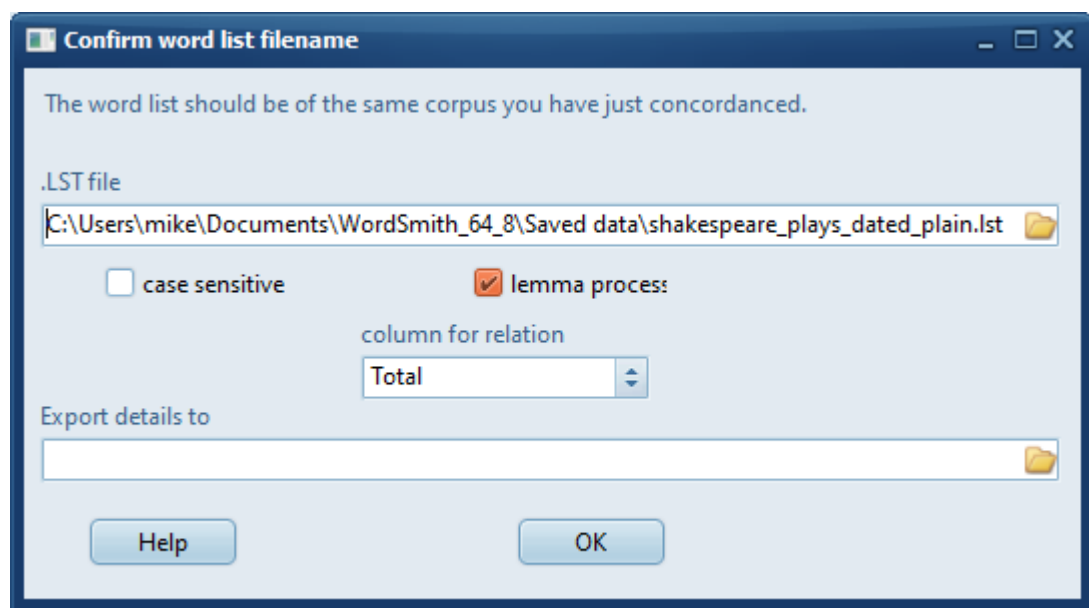
There are nearly 190 instances of **AGO**, mostly in the centre but you see 3 in position L2 (two spaces left of the centre search-word) and 3 more two places to the right of **KISS**. **AND** is the top collocate, found 48 times near **KISS**. Next comes **THE**, usually found to the right of **KISS**. And next **HAND** (not lips!)

However, there are clearly words more grammatical than lexical in the list: grammatical words are usually top frequency items in any language. what's needed is a way of knowing how closely each of these collocates of **KISS** is related to it. Are **AND**, **THE**, **TO** etc. really closely linked to **KISS**?

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 *M/3* column:

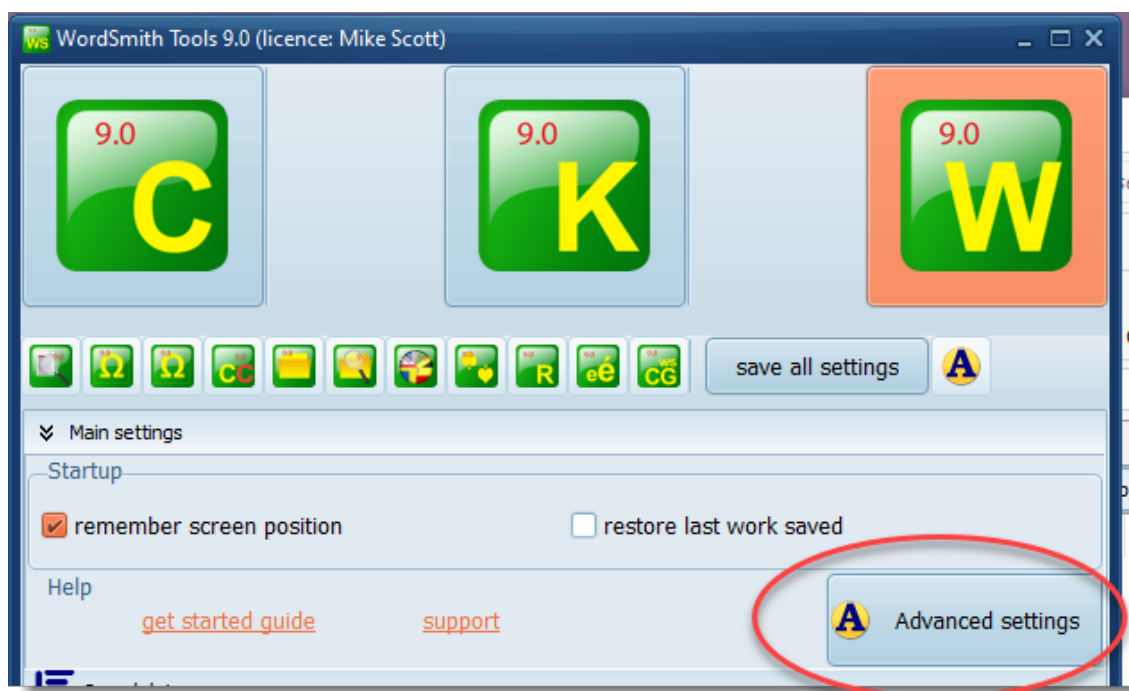
N	Word	With	Dice	MI	M/3	Z	Log_L	T	Log_R	ondProb	ondProb	DeltaPA	DeltaPB
1	KISS	kiss	1.03	12.16	27.21	294.47	248.76	13.78	0.00	103.26	103.26	103.26	103.26
2	HAND	kiss	0.05	6.96	16.01	5.59	74.69	4.76	-2.15	12.50	2.81	12.40	2.79
3	AND	kiss	0.00	3.14	14.31	0.87	0.76	6.14	-7.03	26.09	0.20	23.05	0.18
4	LIPS	kiss	0.05	8.08	14.08	4.18	37.92	2.82	0.49	4.35	6.11	4.33	6.09
5	LOVING	kiss	0.03	7.64	12.29	9.50	20.69	2.22	0.73	2.72	4.50	2.70	4.48
6	THY	kiss	0.01	4.22	12.22	2.54	5.17	3.79	-4.37	8.70	0.42	8.23	0.40
7	THE	kiss	0.00	2.31	12.13	3.98	19.28	4.37	-7.19	16.30	0.11	12.91	0.09
8	ME	kiss	0.00	3.28	11.46	0.12	0.01	3.70	-5.40	9.24	0.22	8.28	0.20
9	ROD	kiss	0.02	9.42	11.42	1.51	13.28	1.41	3.82	1.09	15.38	1.09	15.36
10	YOUR	kiss	0.00	3.39	11.39	0.20	0.04	3.62	-5.20	8.70	0.24	7.86	0.22
11	KATE	kiss	0.03	7.36	11.36	7.62	15.04	1.99	0.77	2.17	3.70	2.16	3.68

The top items in the list now reflect much better the role of a KISS in Shakespeare. (Note: *kiss the rod* meant accept punishment submissively.)

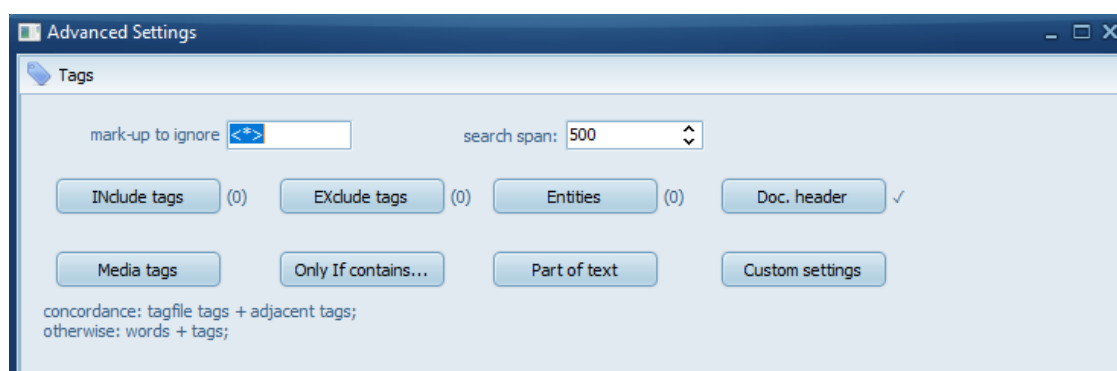
## 4.5 concordancing tagged text (1)

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 *Advanced Settings*



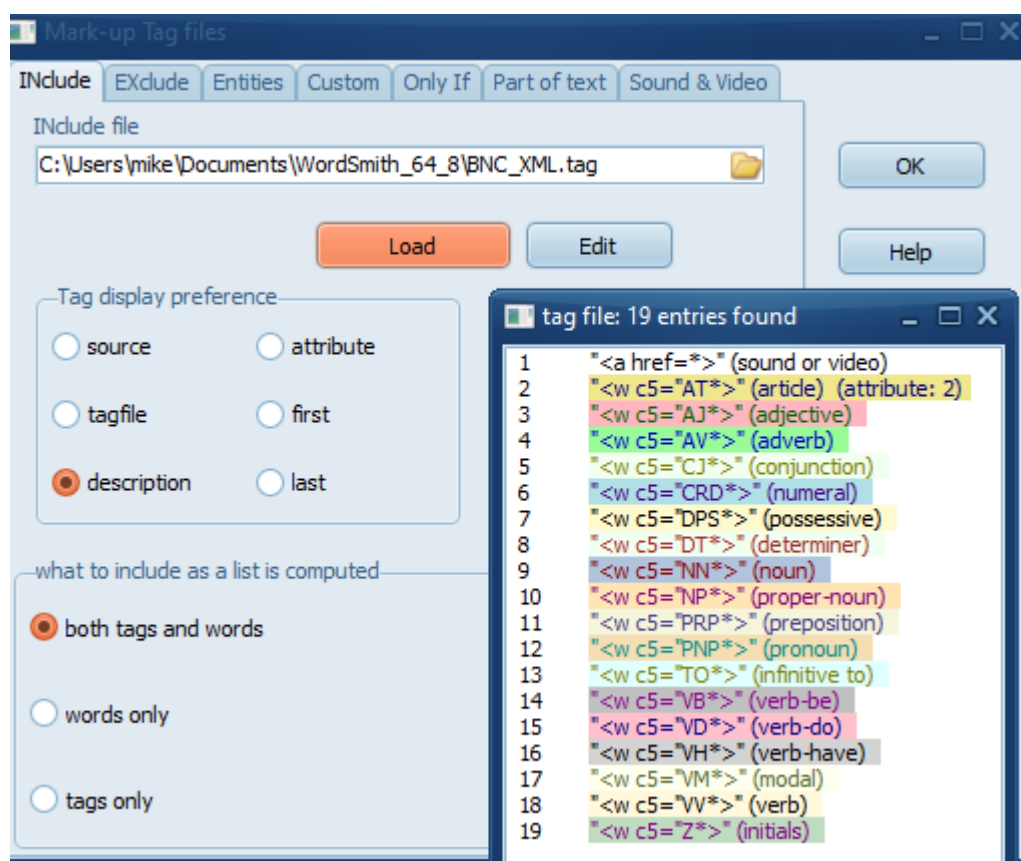


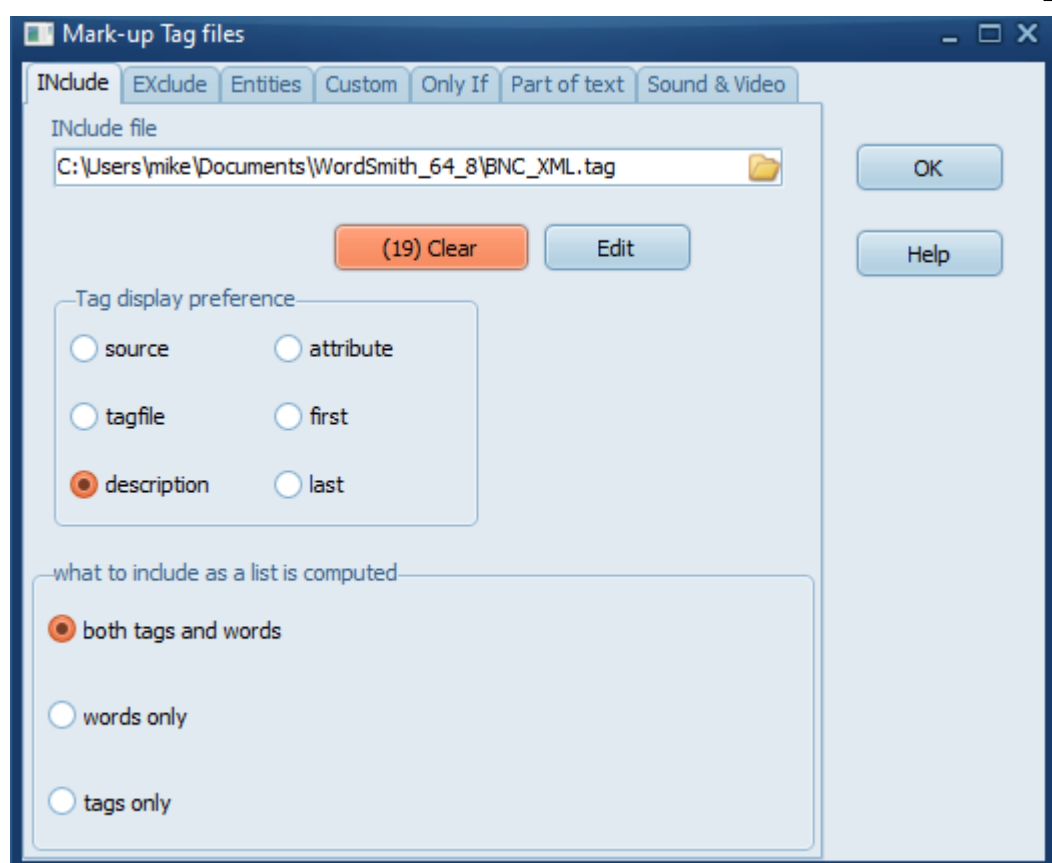
then *Tags*.



By default, WordSmith will ignore anything between < and > brackets if within 500 characters.

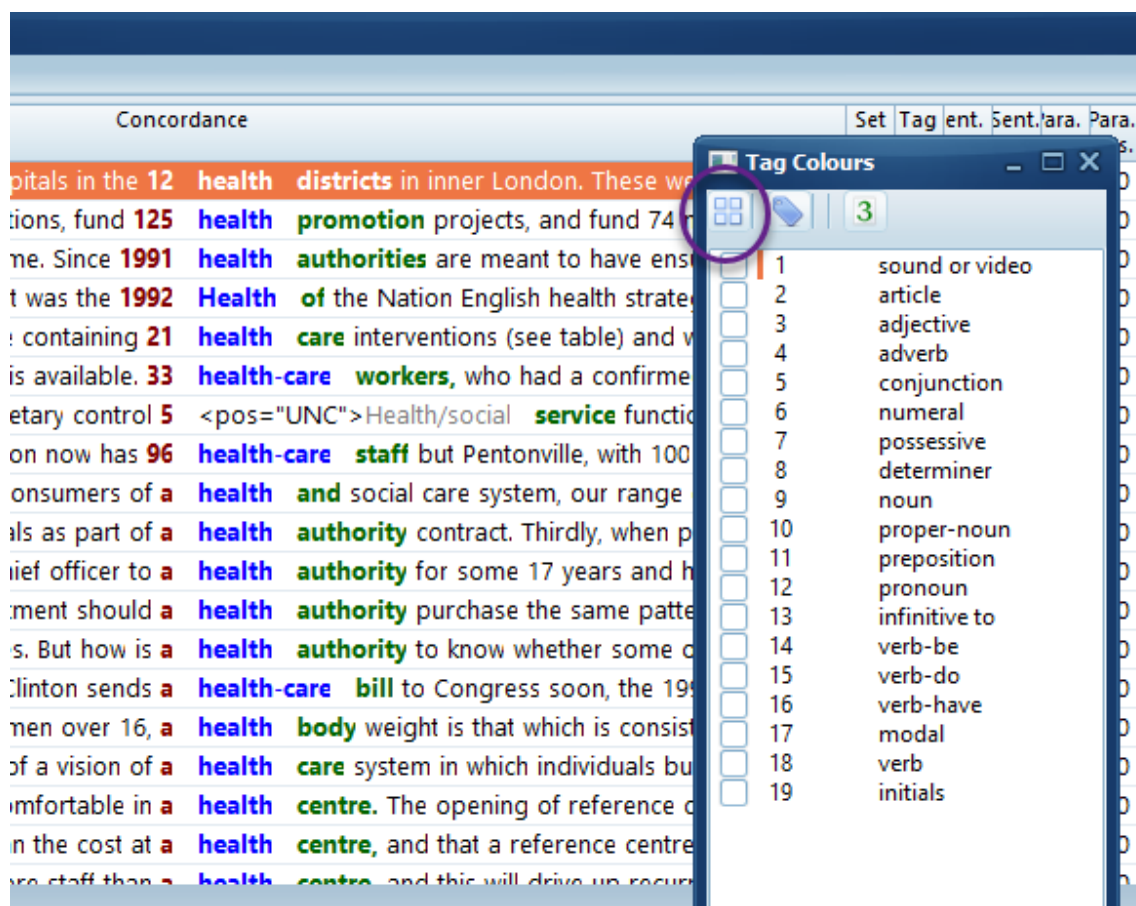
But we may want to use tagged information, if using a tagged corpus like the BNC XML edition. We want to choose the *INclude tags* button so those tags get handled, not ignored.





Press OK. We have 19 different tags which get handled.

If we choose for example medical texts in the BNC and get a concordance on `health`, we might see



Clicking the button in the circle, you get to see the tags as well as the text:

<pos="PREP">in	<pos="ART">the	<pos="ADJ">12	<pos="SUBST">health	<pos="SUBST">districts	
organisations,	<pos="SUBST">fund	<pos="ADJ">125	<pos="SUBST">health	<pos="SUBST">promotion	
programme,	<pos="PREP">Since	<pos="ADJ">1991	<pos="SUBST">health	<pos="SUBST">authorities	
<pos="VERB">was	<pos="ART">the	<pos="ADJ">1992	<pos="SUBST">Health	<pos="ART">of	

or just the words tagged in the colours of their parts of speech:

Concordance					Set	Tag
s working in the 19 general hospitals in the 12 health districts in inner London. These were compared					T">	
orting and 134 cultural organisations, fund 125 health promotion projects, and fund 74 medical,					T">	
t result of the closure programme. Since 1991 health authorities are meant to have ensured that					T">	
missed opportunity on this front was the 1992 Health of the Nation English health strategy, issued by					T">	

and the list of tags lets you switch some on or off:

Concordance				Set	Tag	ent.	Sent.
							Pos.
19 general hospitals in the	12	health districts in inner	London. These were		T">	0	0
tural organisations, fund	125	health promotion projects, and	fund 74 medical,		T">	0	0
sure programme. Since	1991	health authorities are meant to	have ensured that		T">	0	0
ity on this front was the	1992	Health of the Nation English	health strate				
resents a table containing	21	health care interventions (see	table) and				
ture from her is available.	33	health-care workers, who had	a confirme				
egree of budgetary control	5	<pos="UNC">Health/social service					
London. Brixton now has	96	health-care staff but	Pentonville, with 100				
tal illness. As consumers of	a	health and social care system,	our range				
private hospitals as part of	a	health authority contract.	Thirdly, when p				
having been chief officer to	a	health authority for some	17 years and				
is private treatment should	a	health authority purchase	the same patt				
ate inequalities. But how is	a	health authority to know	whether some				
louse. Even if Clinton sends	a	health-care bill to Congress	soon, the 19				
gment. For women over	16,	a health body weight	is that which is				
nces are part of a vision of	a	health care system in	which individuals				

Tag Colours		
1	sound or video	
2	article	
3	adjective	
4	adverb	
5	conjunction	
6	numeral	
7	possessive	
8	determiner	
9	noun	
10	proper-noun	
11	preposition	
12	pronoun	
13	infinitive to	

## 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 c5="NN1"...>
```

to signal singular count nouns. If you want to get all the singular count nouns put

```
<w c5="NN1"*>*
```

as your search word.

Results should be like this:

The screenshot shows the WordSmith Concordance window. The menu bar includes File, Edit, View, Compute, Settings, Windows, and Help. The concordance table has columns for N (line number), Concordance (text), Set, Tag, ent. (frequency), and Sen. Pos. (sentence position). The text in the concordance is from a medical document and includes words like 'abacus', 'abandonment', 'abatement', 'abattoir', 'abbreviation', and 'abdomen'. The word 'abdomen' is highlighted in blue in the original image. Below the table, there are tabs for concordance, collocates, plot, patterns, clusters, timeline, filenames, and notes. At the bottom, there is a status bar showing the current line (108) and total hits (273902), along with buttons for set, navigation, and help.

N	Concordance	Set	Tag	ent.	Sen. Pos.
94	pedigree-drawing program that runs, not on an abacus, but on Windows, the computer	T">		0	0
95	the behaviour are actual or threatened loss or abandonment, or an impasse in a personal	T">		0	0
96	than Vesalius, Paracelsus called for the abandonment of galenism and its	T">		0	0
97	approach may avoid the extremes of complete abandonment of dietary principles on the	T">		0	0
98	junior doctors we cannot condone the unilateral abandonment of agreed manpower targets,	T">		0	0
99	(similar to those operating in the NHS). Pension abatement has been singled out as a blatant	T">		0	0
100	Local Government Superannuation believes that abatement rules are 'illogical and	T">		0	0
101	taxation and superannuation rules (especially abatement provisions), and employment	T">		0	0
102	Bovine gall bladder was obtained from the local abattoir and transferred on ice to the	T">		0	0
103	These must be written legibly and without abbreviation, so that the student can make	T">		0	0
104	or misinterpreted because of illegibility or abbreviation. The ward sister and doctor	T">		0	0
105	comments on. In one of my papers I used an abbreviation in a non-standard way, partly	T">		0	0
106	hydrate. Three hours after the ileal infusion, the abdomen was opened and the mesenteric	T">		0	0
107	obstruction. In all patients, plain films of the abdomen showed abnormal gastrointestinal	T">		0	0
108	have undergone extensive surgery in the upper abdomen. Problems may also be	T">		0	0
109	animals were fasted for 24 hours and then their abdomen was opened and the stomach	T">		0	0
110	She was seen to be afebrile with a tender rigid abdomen on admission to her local hospital.	T">		0	0
111	presenting with pain in the lower chest or upper abdomen. Pain arising from the costal	T">		0	0
112	rats were killed by cervical dislocation and the abdomen was opened immediately. The	T">		0	0
113	entry in the right lower chest and a distended abdomen with tenderness and guarding,	T">		0	0

and if you choose to show the words coloured:

This screenshot shows the same concordance table as above, but with words colored based on their frequency. A vertical color bar on the right side of the table indicates the frequency of each word, with a scale from 1 to 17. The colors range from red (low frequency) to yellow (high frequency). The word 'abdomen' is highlighted in blue, and its frequency is indicated by a yellow bar on the right.

N	Concordance	Set	Tag	ent.	Sen. Pos.
94	pedigree-drawing program that runs, not on an abacus, but on Windows, the computer	T">		0	0
95	the behaviour are actual or threatened loss or abandonment, or an impasse in a personal	T">		0	0
96	than Vesalius, Paracelsus called for the abandonment of galenism and its	T">		0	0
97	approach may avoid the extremes of complete abandonment of dietary principles on the	T">		0	0
98	junior doctors we cannot condone the unilateral abandonment of agreed manpower targets,	T">		0	0
99	(similar to those operating in the NHS). Pension abatement has been singled out as a blatant	T">		0	0
100	Local Government Superannuation believes that abatement rules are 'illogical and	T">		0	0
101	taxation and superannuation rules (especially abatement provisions), and employment	T">		0	0
102	Bovine gall bladder was obtained from the local abattoir and transferred on ice to the	T">		0	0
103	These must be written legibly and without abbreviation, so that the student can make	T">		0	0
104	or misinterpreted because of illegibility or abbreviation. The ward sister and doctor may	T">		0	0
105	comments on. In one of my papers I used an abbreviation in a non-standard way, partly	T">		0	0
106	hydrate. Three hours after the ileal infusion, the abdomen was opened and the mesenteric	T">		0	0
107	obstruction. In all patients, plain films of the abdomen showed abnormal gastrointestinal	T">		0	0
108	have undergone extensive surgery in the upper abdomen. Problems may also be	T">		0	0
109	animals were fasted for 24 hours and then their abdomen was opened and the stomach	T">		0	0
110	She was seen to be afebrile with a tender rigid abdomen on admission to her local hospital.	T">		0	0
111	presenting with pain in the lower chest or upper abdomen. Pain arising from the costal	T">		0	0
112	rats were killed by cervical dislocation and the abdomen was opened immediately. The	T">		0	0
113	entry in the right lower chest and a distended abdomen with tenderness and guarding,	T">		0	0

## Number of hits

Note that we got 273,902 hits. That is a lot of concordance entries! Nouns are very common especially in written medical English. To speed things up I altered a setting in the main Controller:

Concord settings

What you get What you see

Search settings

Cluster settings

context to save (words)

100

Collocates

left span right span

L5 R5

☐ case sensitive

save-as-text search-word marker

save-as-text context-word marker

☒ concordance only

min. freq. min. length min. texts

2 2 1

stop at

stop at sentence break

☒ separate search-words

which saved a lot of time in working out the collocates, clusters and plot information and got to show results with less delay. You could press the button with the red exclamation mark ( ! ) to stop processing.

*Step-by-step guide to WordSmith*

# *WordList*

**Section**

---

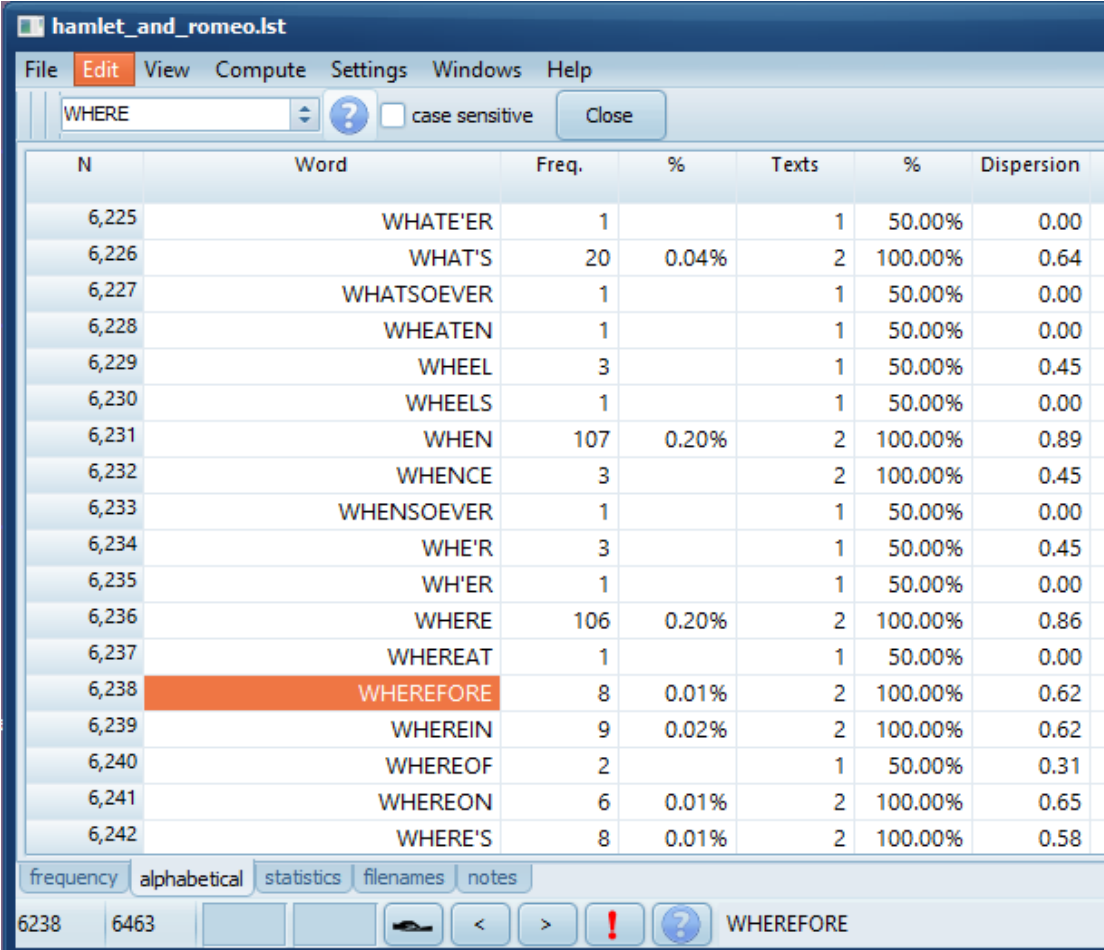




## 5 WordList

### 5.1 overview

A word list in WordSmith Tools looks something like this:



N	Word	Freq.	%	Texts	%	Dispersion
6,225	WHATE'ER	1		1	50.00%	0.00
6,226	WHAT'S	20	0.04%	2	100.00%	0.64
6,227	WHATSOEVER	1		1	50.00%	0.00
6,228	WHEATEN	1		1	50.00%	0.00
6,229	WHEEL	3		1	50.00%	0.45
6,230	WHEELS	1		1	50.00%	0.00
6,231	WHEN	107	0.20%	2	100.00%	0.89
6,232	WHENCE	3		2	100.00%	0.45
6,233	WHENSOEVER	1		1	50.00%	0.00
6,234	WHE'R	3		1	50.00%	0.45
6,235	WH'ER	1		1	50.00%	0.00
6,236	WHERE	106	0.20%	2	100.00%	0.86
6,237	WHEREAT	1		1	50.00%	0.00
6,238	WHEREFORE	8	0.01%	2	100.00%	0.62
6,239	WHEREIN	9	0.02%	2	100.00%	0.62
6,240	WHEREOF	2		1	50.00%	0.31
6,241	WHEREON	6	0.01%	2	100.00%	0.65
6,242	WHERE'S	8	0.01%	2	100.00%	0.58

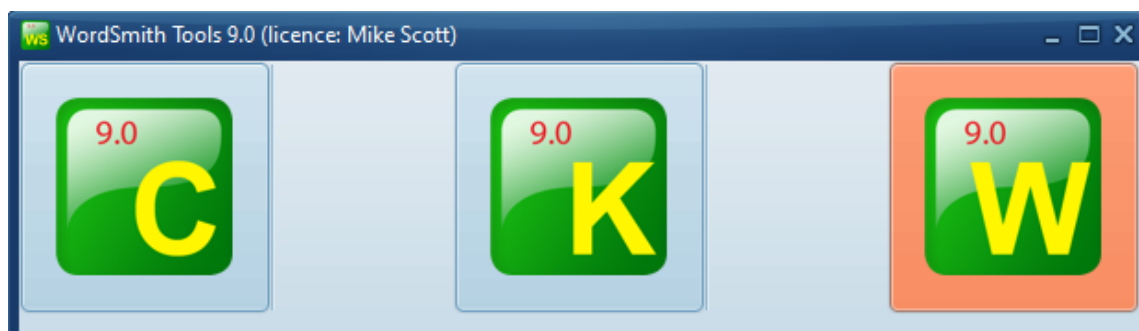
frequency alphabetical statistics filenames notes

6238 6463 < > ! ? WHEREFORE

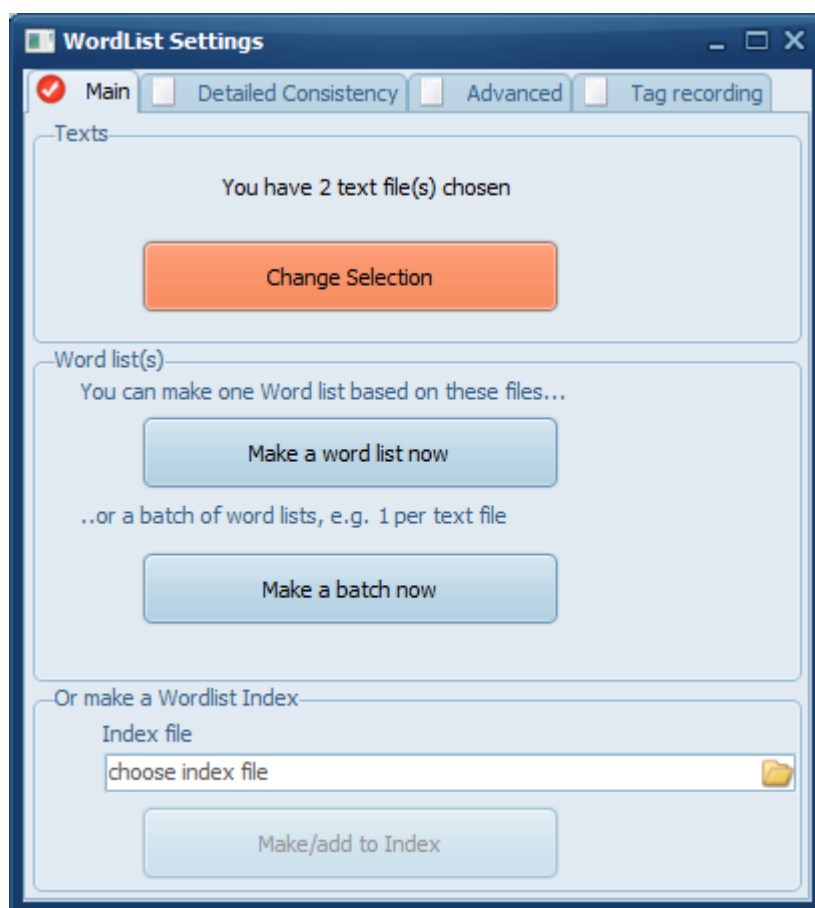
It shows how often each word occurs in the text files, what that is as a percent of the running words in the text, and how many text files each word was found in.

## 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.



Here we're going to make one simple word-list based on 2 text files (the plays *Romeo and Juliet* and *Hamlet*), so press *Make a word list now*.

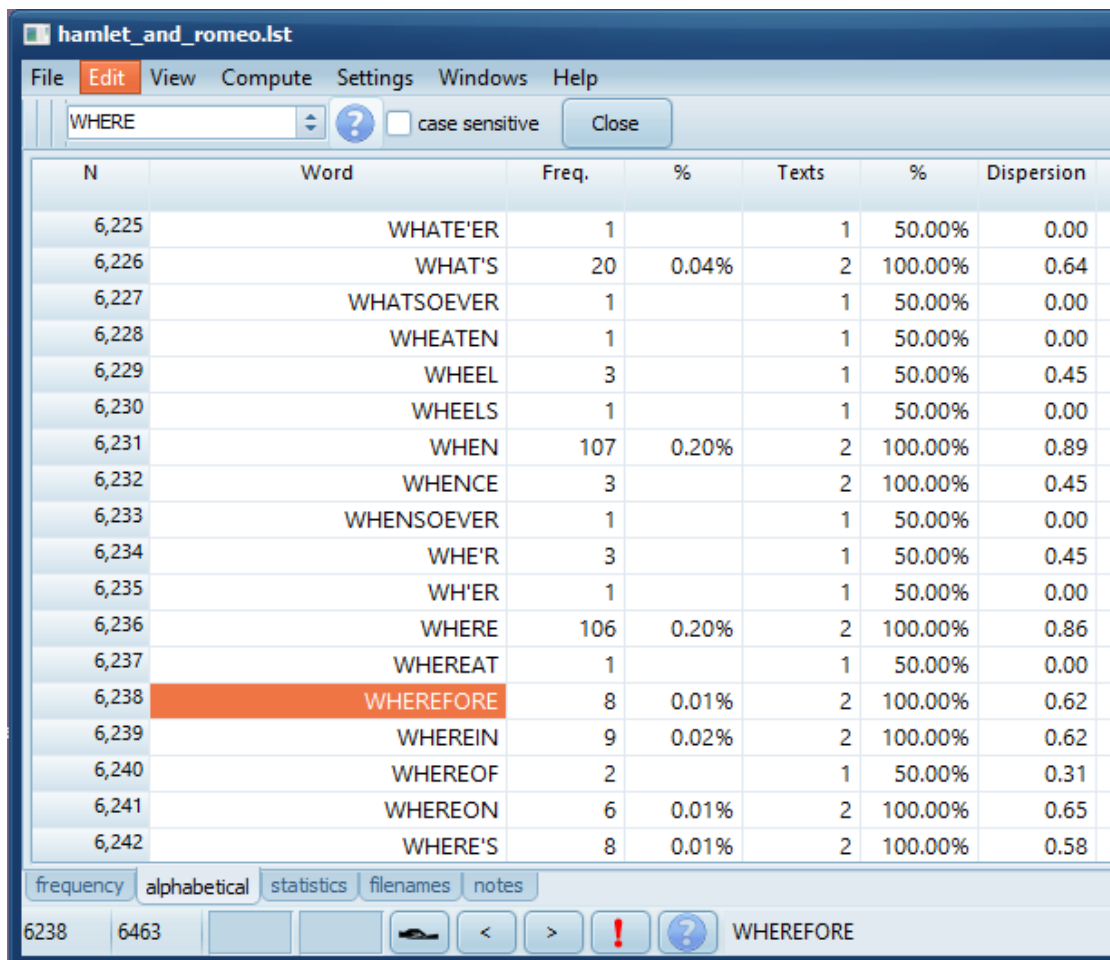
N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
1	THE	1,757	3.26%	2	100.00%	0.95		
2	AND	1,565	2.90%	2	100.00%	0.96		
3	TO	1,298	2.41%	2	100.00%	0.95		
4	I	1,141	2.12%	2	100.00%	0.95		
5	OF	1,044	1.94%	2	100.00%	0.94		
6	A	970	1.80%	2	100.00%	0.92		
7	MY	875	1.62%	2	100.00%	0.94		
8	YOU	846	1.57%	2	100.00%	0.90		
9	IN	742	1.38%	2	100.00%	0.96		
10	THAT	741	1.37%	2	100.00%	0.94		
11	IS	690	1.28%	2	100.00%	0.94		
12	IT	640	1.19%	2	100.00%	0.92		
13	NOT	574	1.06%	2	100.00%	0.95		
14	THIS	512	0.95%	2	100.00%	0.93		
15	ME	500	0.93%	2	100.00%	0.92		
16	WITH	494	0.92%	2	100.00%	0.93		
17	FOR	474	0.88%	2	100.00%	0.94		
18	BUT	448	0.83%	2	100.00%	0.93		
19	BE	440	0.82%	2	100.00%	0.95		

frequency alphabetical statistics filenames notes

8 6463 THE

The WordList tool shows us a frequency listing. The most frequent words are **THE**, **AND**, **TO**, etc.. Beside each one you can see how frequent it is in the collection of 2 texts we used, the percentage of running words, and how many of our 2 texts each word occurred in.

To see the words in alphabetical order instead, click the alphabetical tab near the bottom of the window.



hamlet\_and\_romeo.lst

File Edit View Compute Settings Windows Help

WHERE ? case sensitive Close

N	Word	Freq.	%	Texts	%	Dispersion
6,225	WHATE'ER	1		1	50.00%	0.00
6,226	WHAT'S	20	0.04%	2	100.00%	0.64
6,227	WHATSOEVER	1		1	50.00%	0.00
6,228	WHEATEN	1		1	50.00%	0.00
6,229	WHEEL	3		1	50.00%	0.45
6,230	WHEELS	1		1	50.00%	0.00
6,231	WHEN	107	0.20%	2	100.00%	0.89
6,232	WHENCE	3		2	100.00%	0.45
6,233	WHENSOEVER	1		1	50.00%	0.00
6,234	WHE'R	3		1	50.00%	0.45
6,235	WH'ER	1		1	50.00%	0.00
6,236	WHERE	106	0.20%	2	100.00%	0.86
6,237	WHEREAT	1		1	50.00%	0.00
6,238	WHEREFORE	8	0.01%	2	100.00%	0.62
6,239	WHEREIN	9	0.02%	2	100.00%	0.62
6,240	WHEREOF	2		1	50.00%	0.31
6,241	WHEREON	6	0.01%	2	100.00%	0.65
6,242	WHERE'S	8	0.01%	2	100.00%	0.58

frequency alphabetical statistics filenames notes

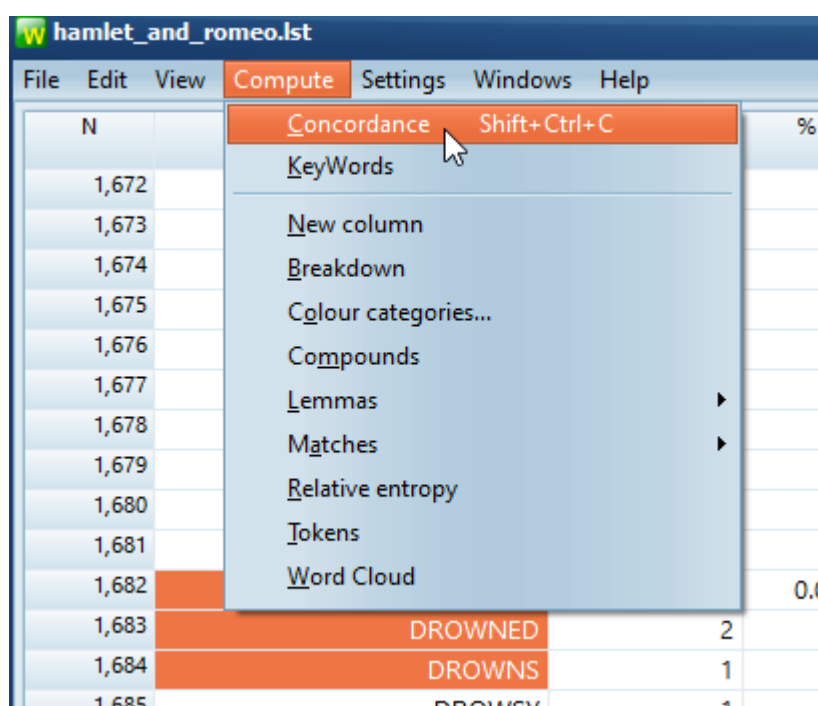
6238 6463 < > ! ? WHEREFORE

Now scroll down to **wherefore**. There are 8 cases of this word, that is 0.01% of the running words of the two plays; it is present in both plays, not just one.

### 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 one 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):

N	Concordance			Se
1	then turn tears to fires! And these, who <b>often</b>	<b>drown'd</b>	could never die, Transparent	
2	another's heel, So fast they follow: your <b>sister's</b>	<b>drown'd,</b>	Laertes. Drown'd! O, where? There	
3	fast they follow: your sister's drown'd, <b>Laertes.</b>	<b>Drown'd!</b>	O, where? There is a willow grows	
4	lay To muddy death. Alas! then, she <b>is</b>	<b>drown'd?</b>	Drown'd, drown'd. Too much of	
5	To muddy death. Alas! then, she <b>is drown'd?</b>	<b>Drown'd,</b>	drown'd. Too much of water hast	
6	death. Alas! then, she is drown'd? <b>Drown'd,</b>	<b>drown'd.</b>	Too much of water hast thou, poor	
7	it Christian burial. How can that be, unless <b>she</b>	<b>drowned</b>	herself in her own defence? Why,	
8	it is, to act, to do, and to perform: argal, <b>she</b>	<b>drowned</b>	herself wittingly. Nay, but hear you,	
9	if the water come to him, and drown him, <b>he</b>	<b>drowns</b>	not himself: argal, he that is not guilty	

## 5.4 lemmatising

To lemmatise manually, with a word list on screen,

DROSSY	1		1
DROWN	5		1
DROWN'D	6	0.01%	2
DROWNED	2		1
DROWNS	1		1

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

and drop it:

1,680	DROSSY	1		1	50.00%	0.00	
1,681	DROWN	11	0.02%	n/a	n/a	n/a	DROWN[5] DROWN'D[6]
1,682	DROWN'D	6	0.01%	2	100.00%	0.16	
1,683	DROWNED	2		1			
1,684	DROWNS	1		1			

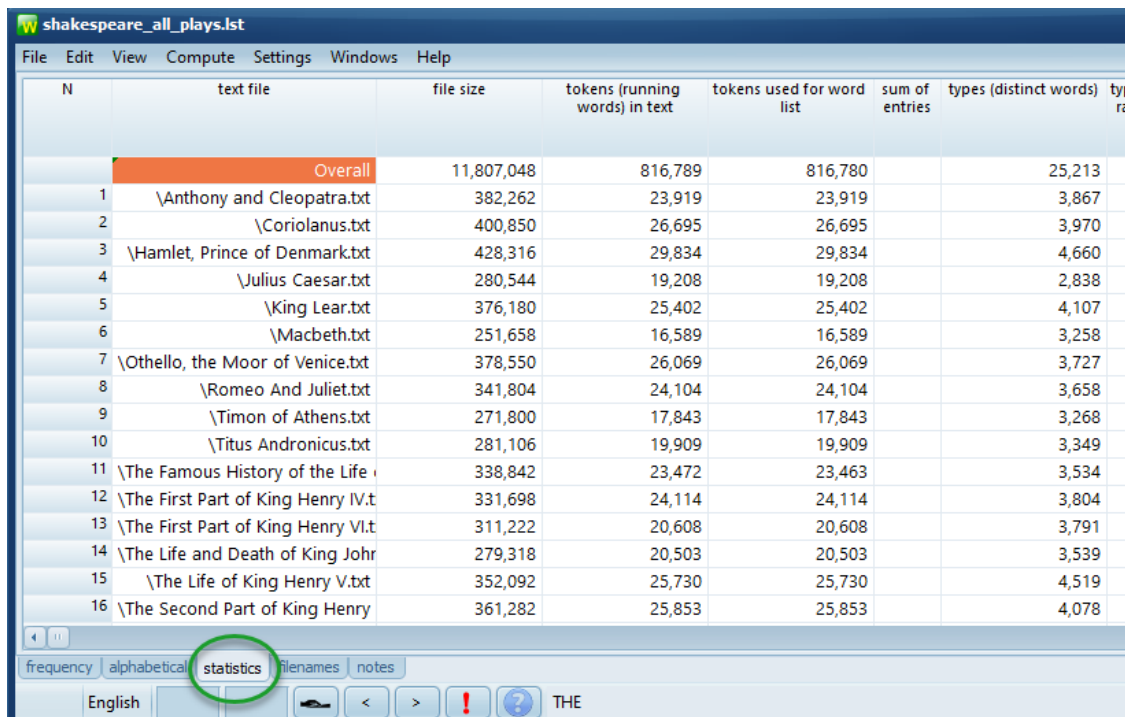
You will then see the totals change and the items become visible in the Lemmas column.

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

Lemma Forms	
Variants	Frequency
DROWN	5
DROWN'D	6
DROWNED	2
DROWNS	1

## 5.5 word list statistics

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



N	text file	file size	tokens (running words) in text	tokens used for word list	sum of entries	types (distinct words)	types (distinct words) in word list
	<b>Overall</b>	11,807,048	816,789	816,780		25,213	
1	\Anthony and Cleopatra.txt	382,262	23,919	23,919		3,867	
2	\Coriolanus.txt	400,850	26,695	26,695		3,970	
3	\Hamlet, Prince of Denmark.txt	428,316	29,834	29,834		4,660	
4	\Julius Caesar.txt	280,544	19,208	19,208		2,838	
5	\King Lear.txt	376,180	25,402	25,402		4,107	
6	\Macbeth.txt	251,658	16,589	16,589		3,258	
7	\Othello, the Moor of Venice.txt	378,550	26,069	26,069		3,727	
8	\Romeo And Juliet.txt	341,804	24,104	24,104		3,658	
9	\Timon of Athens.txt	271,800	17,843	17,843		3,268	
10	\Titus Andronicus.txt	281,106	19,909	19,909		3,349	
11	\The Famous History of the Life of King Henry the Fourth.txt	338,842	23,472	23,463		3,534	
12	\The First Part of King Henry IV.txt	331,698	24,114	24,114		3,804	
13	\The First Part of King Henry VI.txt	311,222	20,608	20,608		3,791	
14	\The Life and Death of King John.txt	279,318	20,503	20,503		3,539	
15	\The Life of King Henry V.txt	352,092	25,730	25,730		4,519	
16	\The Second Part of King Henry VI.txt	361,282	25,853	25,853		4,078	

and something like this should appear. Lots of numbers.

## 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.	%	Texts	
1	PART OF THE	16,126	0.02%	2,565	81
2	END OF THE	12,248	0.01%	2,424	77
3	MEMBERS OF THE	7,662		1,765	56
4	SIDE OF THE	6,099		1,638	52
5	LIKELY TO BE	5,508		1,545	49
6	SECRETARY OF STATE	4,975		564	17
7	REST OF THE	4,800		1,733	55
8	CENT OF THE	4,755		923	29
9	MEMBER OF THE	4,630		1,528	48
10	ACCORDING TO THE	4,567		1,543	49
11	NEED TO BE	4,054		1,333	42
12	TOP OF THE	3,883		1,319	41
13	PARTS OF THE	3,828		1,440	45
14	PART OF A	3,690		1,548	49
15	NATURE OF THE	3,568		1,175	37
16	USE OF THE	3,257		1,239	39
17	VIEW OF THE	3,175		1,320	42
18	BASED ON THE	3,156		1,278	40
19	WAY IN WHICH	3,153		1,036	32
20	LOOK AT THE	3,057		1,060	35

frequency alphabetical statistics filenames notes

1 252866 < > ! ? PART OF THE

Press Ctrl/F2 to save it, and the suggested file-name will be something like `_index_3-5-word clusters`. It can later be opened as an ordinary word-list.



*Step-by-step guide to WordSmith*

# *KeyWords*

**Section**

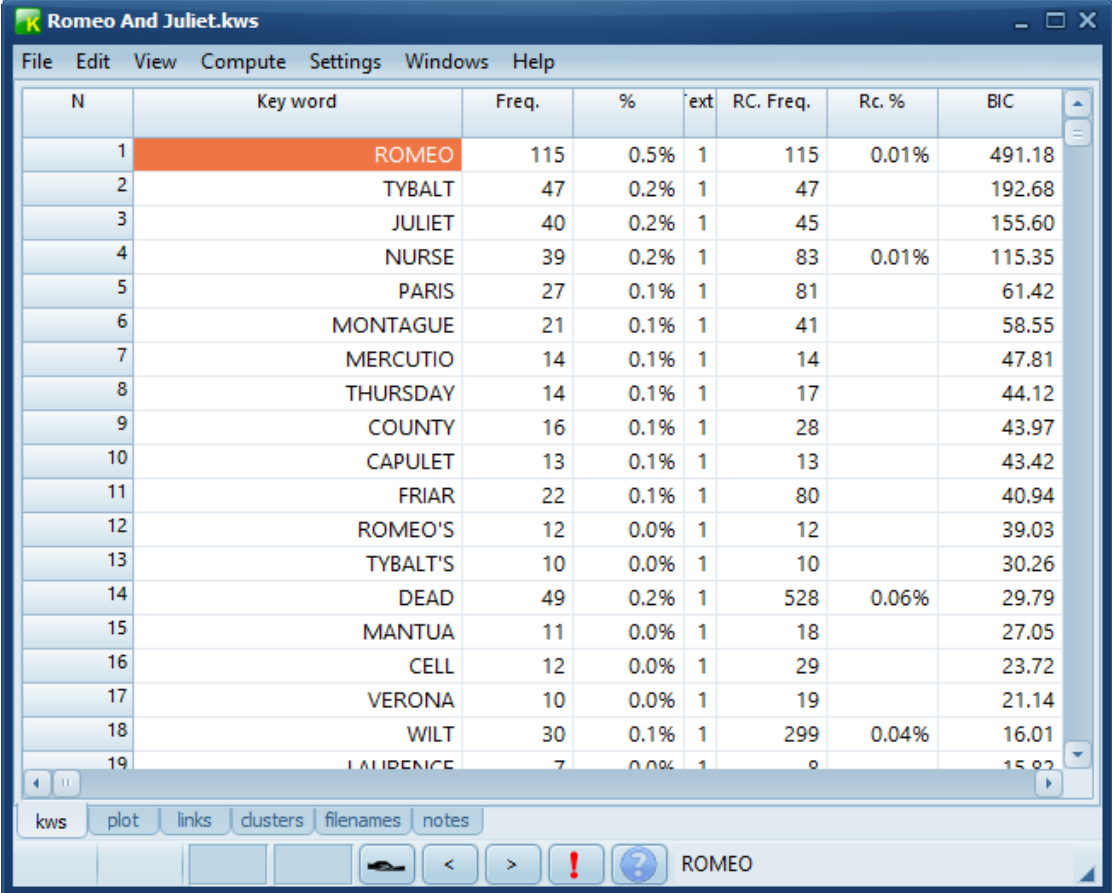
---

**VI**

## 6 KeyWords

### 6.1 overview

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



N	Key word	Freq.	%	ext	RC. Freq.	Rc. %	BIC
1	ROMEO	115	0.5%	1	115	0.01%	491.18
2	TYBALT	47	0.2%	1	47		192.68
3	JULIET	40	0.2%	1	45		155.60
4	NURSE	39	0.2%	1	83	0.01%	115.35
5	PARIS	27	0.1%	1	81		61.42
6	MONTAGUE	21	0.1%	1	41		58.55
7	MERCUTIO	14	0.1%	1	14		47.81
8	THURSDAY	14	0.1%	1	17		44.12
9	COUNTY	16	0.1%	1	28		43.97
10	CAPULET	13	0.1%	1	13		43.42
11	FRIAR	22	0.1%	1	80		40.94
12	ROMEO'S	12	0.0%	1	12		39.03
13	TYBALT'S	10	0.0%	1	10		30.26
14	DEAD	49	0.2%	1	528	0.06%	29.79
15	MANTUA	11	0.0%	1	18		27.05
16	CELL	12	0.0%	1	29		23.72
17	VERONA	10	0.0%	1	19		21.14
18	WILT	30	0.1%	1	299	0.04%	16.01
19	LAURENCE	7	0.0%	1	8		15.82

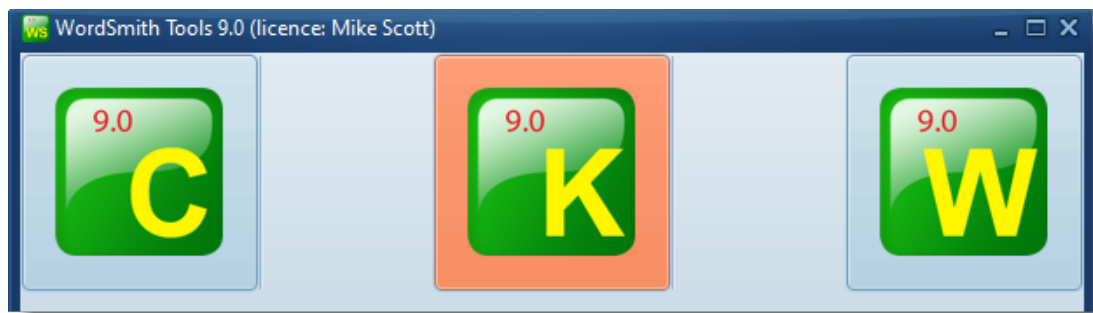
The key words are words which occur unusually frequently in comparison with some kind of reference corpus.

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

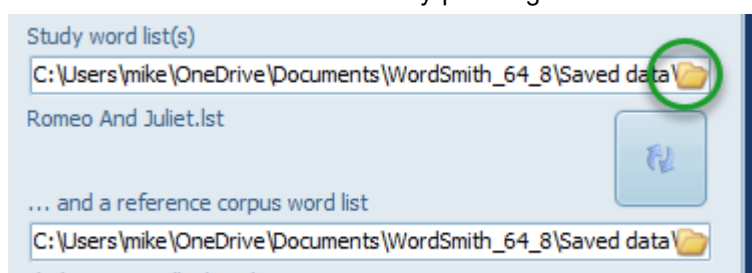
To make a key word list, first press the KeyWords button in the main Controller.



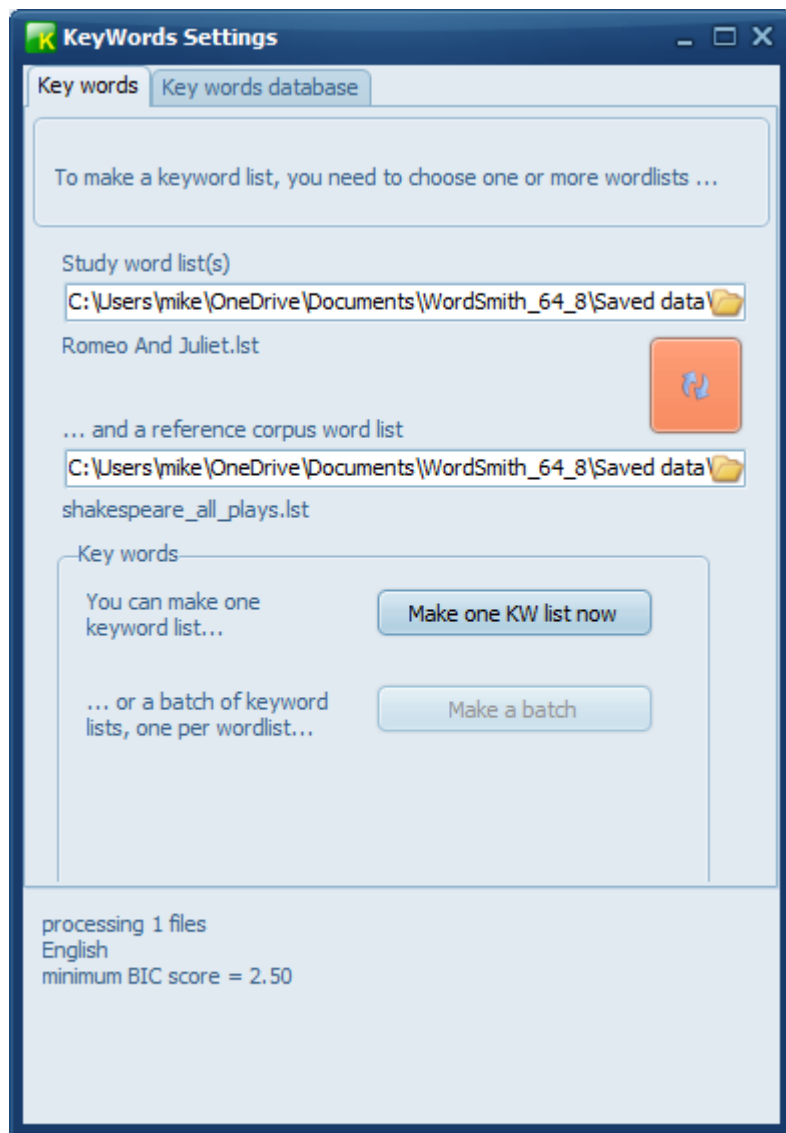
When KeyWords starts up, choose menu option *File*, then *New* and you will see something like this.



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



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

Romeo And Juliet.kws

File Edit View Compute Settings Windows Help

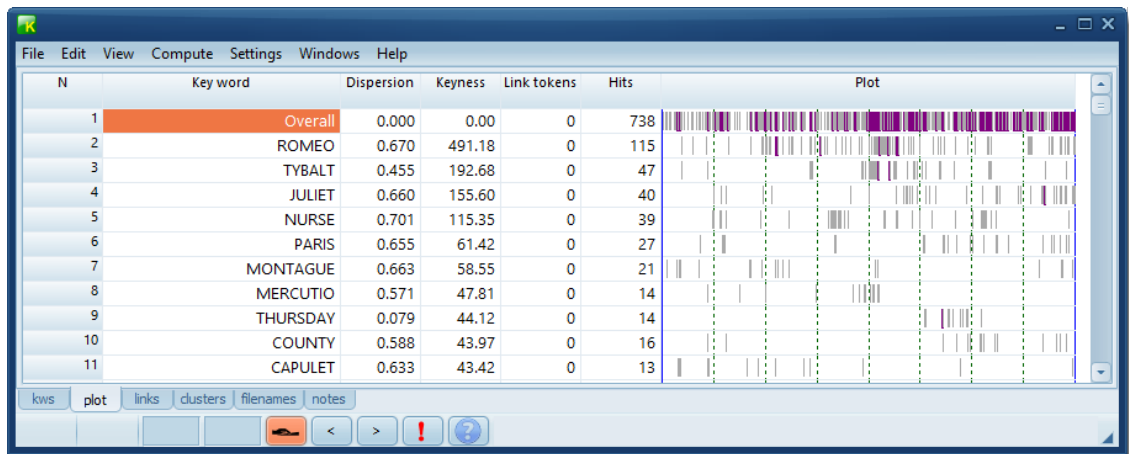
N	Key word	Freq.	%	ext	RC. Freq.	Rc. %	BIC
1	ROMEO	115	0.5%	1	115	0.01%	491.18
2	TYBALT	47	0.2%	1	47		192.68
3	JULIET	40	0.2%	1	45		155.60
4	NURSE	39	0.2%	1	83	0.01%	115.35
5	PARIS	27	0.1%	1	81		61.42
6	MONTAGUE	21	0.1%	1	41		58.55
7	MERCUTIO	14	0.1%	1	14		47.81
8	THURSDAY	14	0.1%	1	17		44.12
9	COUNTY	16	0.1%	1	28		43.97
10	CAPULET	13	0.1%	1	13		43.42
11	FRIAR	22	0.1%	1	80		40.94
12	ROMEO'S	12	0.0%	1	12		39.03
13	TYBALT'S	10	0.0%	1	10		30.26
14	DEAD	49	0.2%	1	528	0.06%	29.79
15	MANTUA	11	0.0%	1	18		27.05
16	CELL	12	0.0%	1	29		23.72
17	VERONA	10	0.0%	1	19		21.14
18	WILT	30	0.1%	1	299	0.04%	16.01
19	LAURENCE	7	0.0%	1	9		15.82

kws plot links clusters filenames notes

ROMEO

### 6.3 key words plot

This is a key word plot where the text is Romeo and Juliet, compared with all of the Shakespeare plays.



You see:

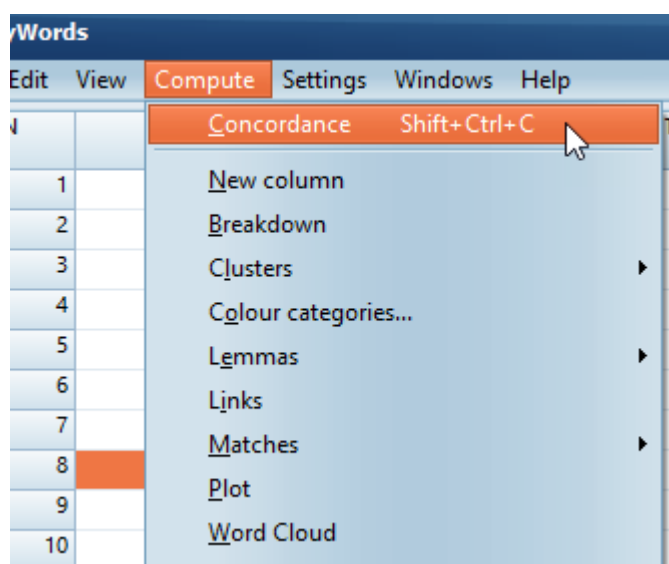
- each key word (KW)
- a map showing where each word came.

The left edge represents the start of the text and the right edge represents the end. Notice how **NURSE** and **TYBALT** seem to come in bursts at various points through the play.

### 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 **THURSDAY** 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
1	ha, ha! Well, Wednesday is too soon; O' <b>Thursday</b> let it be: o' Thursday, tell her, She		
2	is too soon; O' Thursday let it be: o' <b>Thursday</b> , tell her, She shall be married to		
3	friends, And there an end. But what say you <b>to Thursday?</b> My lord, I would that Thursday		
4	say you to Thursday? My lord, I would <b>that Thursday</b> were to-morrow. Well, get you		
5	were to-morrow. Well, get you gone: o' <b>Thursday</b> be it then. Go you to Juliet ere you		
6	what day is that? Marry, my child, early <b>next Thursday</b> morn The gallant, young, and		
7	me no prouds, But fettle your fine joints ' <b>gainst Thursday</b> next, To go with Paris to Saint		
8	wretch! I tell thee what, get thee to church o' <b>Thursday</b> , Or never after look me in the face.		
9	me: Look to't, think on't, I do not use to <b>jest. Thursday</b> is near; lay hand on heart, advise.		
10	If all else fail, myself have power to die. <b>On Thursday</b> , sir? the time is very short. My		
11	I may be a wife. That may be must be, love, <b>on Thursday</b> next. What must be shall be. That's		
12	God shield, I should disturb devotion! Juliet, <b>on Thursday</b> early will I rouse you: Till then,		
13	thou must, and nothing may proroque it, <b>On Thursday</b> next be married to this county. Tell		
14	think fit to furnish me to-morrow? No, not <b>till Thursday</b> : there is time enough. Go, nurse,		



---

# Index

## - C -

choosing text files 5  
collocates and mutual information 19  
Concord: nearest tag 25  
Concord: overview 13  
concordancing on tags 20

## - I -

introduction 2

## - K -

KeyWords: overview 38

## - M -

making a concordance 13

## - N -

nearest tag 25

## - S -

seeing source text 17  
sorting tags 25

## - T -

tag concordancing 20

## - W -

WordList: overview 29

## Step-by-step guide to WordSmith