



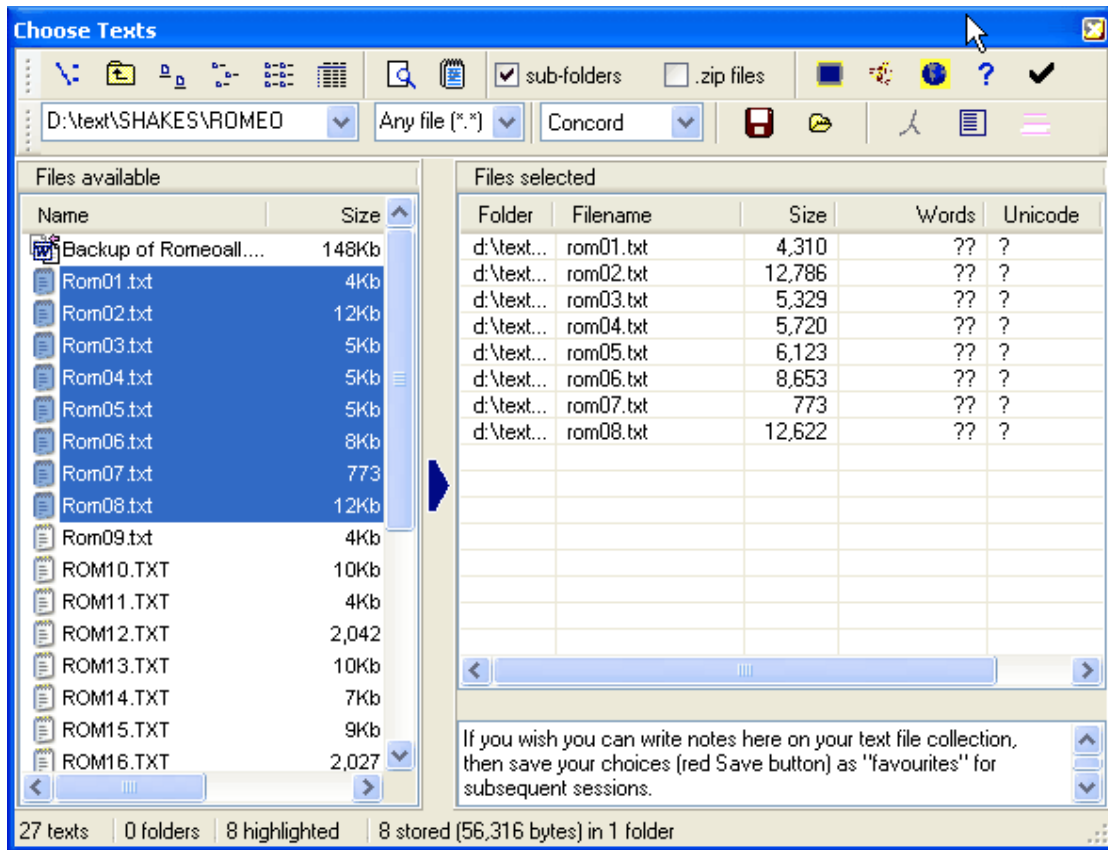
Oxford WordSmith Tools Manual

I. Introduction

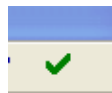
There are four menu-options in the WordSmith tools controller, a continually-changing saying, three buttons for the main tools (the tool in use usually shown in red), and a series of tabs. No texts have been chosen yet for any of the tools:

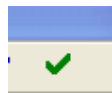


To select files for the *Files selected*-window, click on the big blue arrow, or drag some text files from the *Files available*-window to the right. You should see something like this:



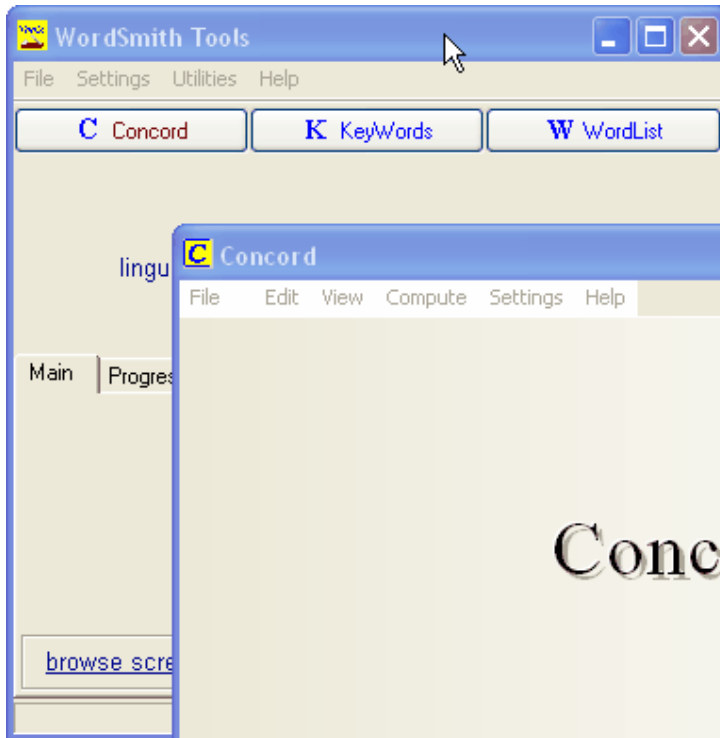
WordSmith shows in the status bar at the bottom that 8 texts have been chosen. The file-sizes are visible, but WordSmith doesn't know yet the number of words contained in each text file.



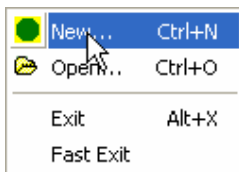
Click on the green tick  or just close the window.

III. Concordancing

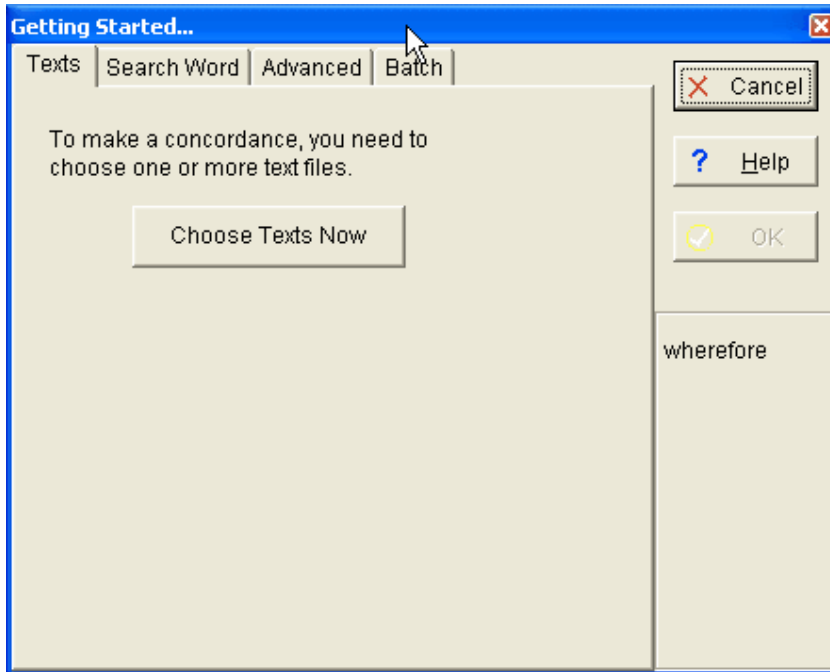
Click on the *Concord*-button and a new window opens:



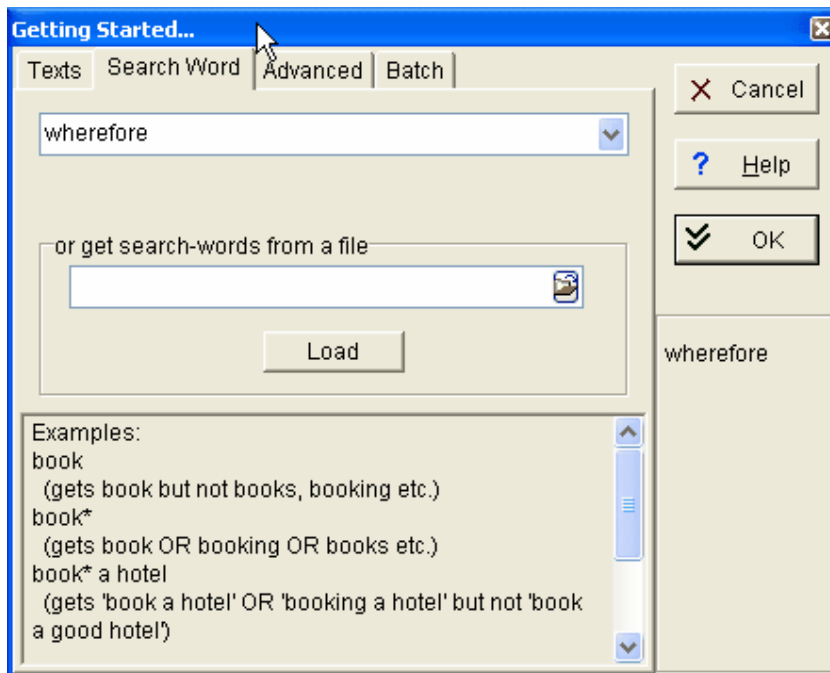
In the new *Concord*-window choose *File | New*:



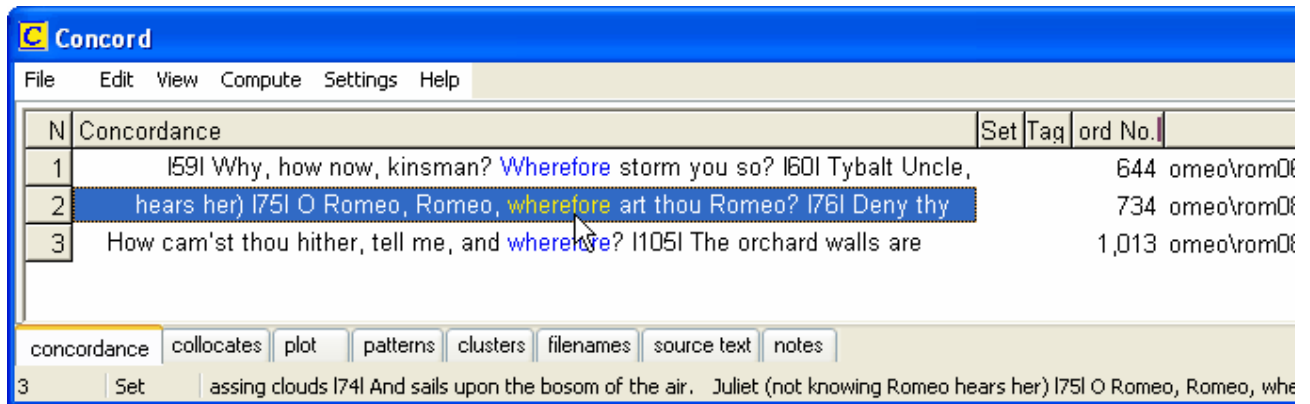
If no text files have been chosen yet, you are asked to choose some.
Click on the *Choose Texts Now* button:



Once the texts have been selected, enter a suitable *Search Word*, for example *wherefore*. Then press *OK*:

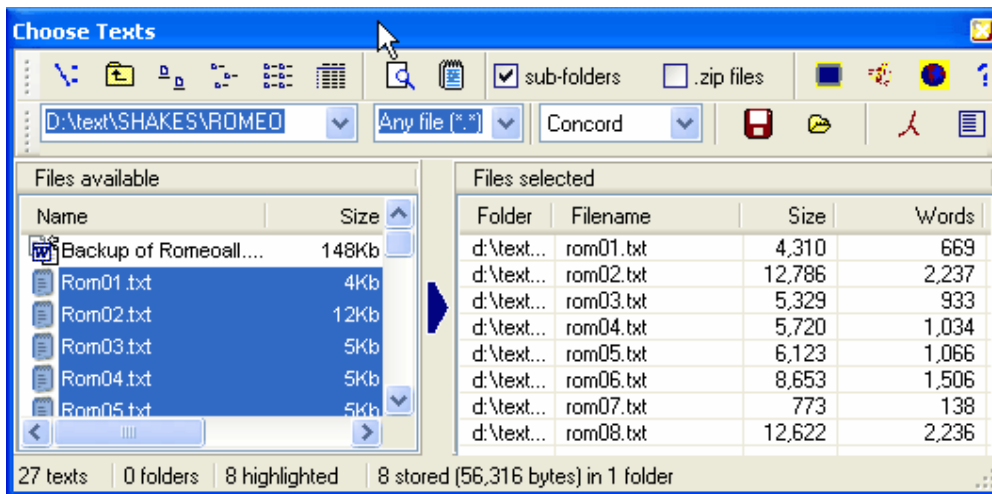


You then get a concordance of all occurrences of *wherefore* in the above-chosen text files:



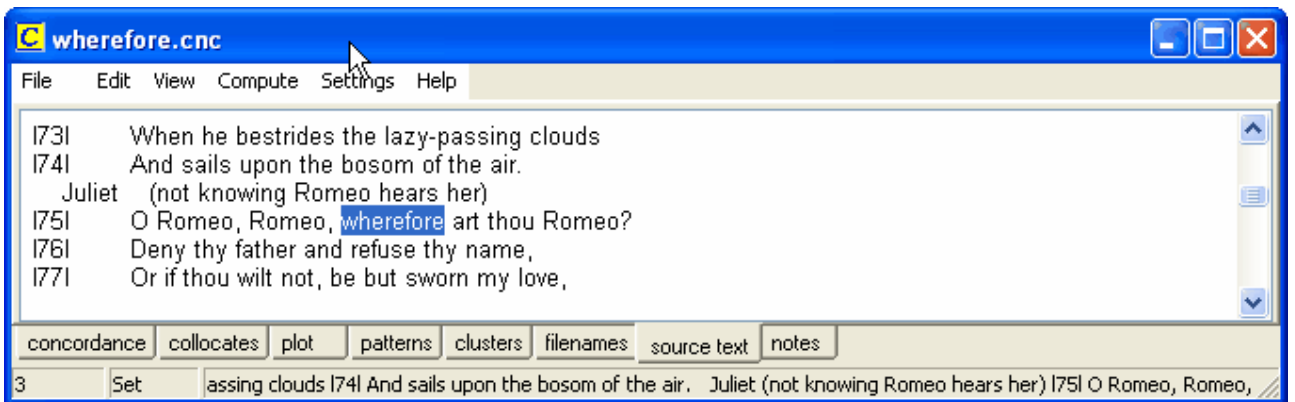
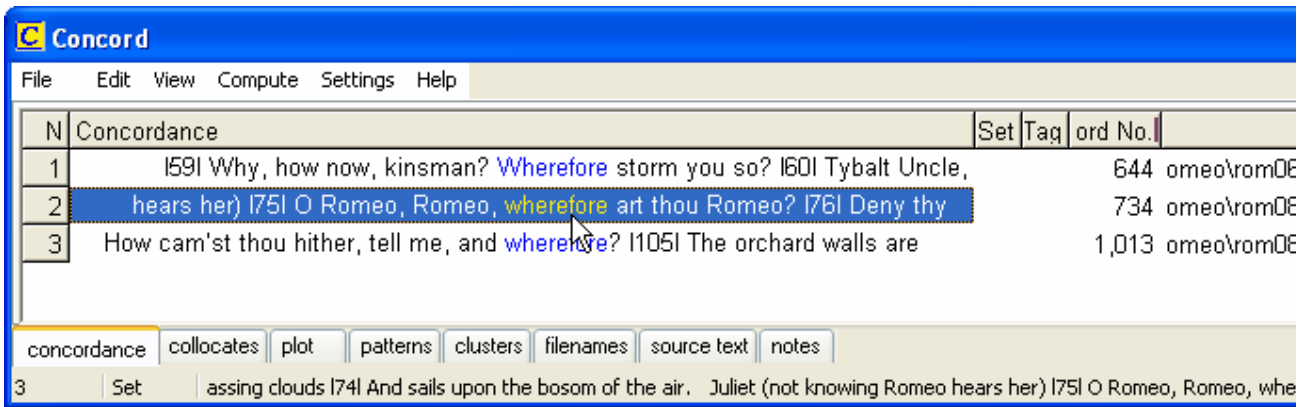
The first instance of *wherefore* occurs as 644th word of one of the chosen texts, *rom06.txt*.

Now, WordSmith also knows the number of words in each text file (e.g. 1,506 words in *rom06.txt*):

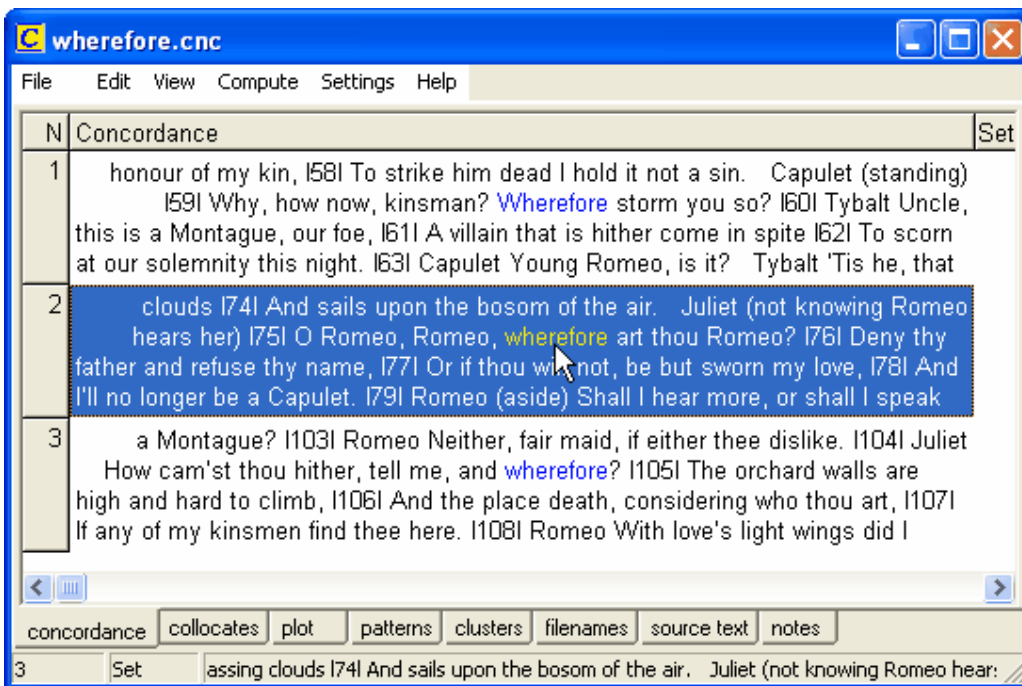


Accessing the Source Text

To access the source text, double-click on the line in question:



or press F8 to expand the context:



or place the cursor between line 2 and 3 in the left column to expand them:

| N | Concordance |
|---|---|
| 1 | 159 Why, how now, kinsman? Wherefore storm you so? 160 Tybalt Uncle |
| 2 | hears her) 175 O Romeo, Romeo, wherefore art thou Romeo? 176 Deny thy |
| 3 | How cam'st thou hither, tell me, and wherefore ? 1105 The orchard walls are |

Collocates and Mutual Information

Here are the collocates of `ago` from the written section of the BNC, ordered by frequency:

The screenshot shows a window titled 'ago.cnc' with a menu bar (File, Edit, View, Compute, Settings, Help) and a table of collocates. The table has columns for rank (N), the word, the relation (all 'ago'), mutual information (all '0.000'), total frequency, and total left frequency. The collocates are ordered by total frequency.

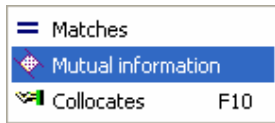
| N | Word | With | elation | Total | tal Left |
|----|--------|------|---------|--------|----------|
| 1 | AGO | ago | 0.000 | 16,785 | 47 |
| 2 | YEARS | ago | 0.000 | 9,033 | 8,936 |
| 3 | A | ago | 0.000 | 6,967 | 4,608 |
| 4 | THE | ago | 0.000 | 6,352 | 1,615 |
| 5 | WAS | ago | 0.000 | 2,951 | 1,183 |
| 6 | OF | ago | 0.000 | 2,949 | 1,345 |
| 7 | AND | ago | 0.000 | 2,740 | 623 |
| 8 | TO | ago | 0.000 | 2,506 | 679 |
| 9 | IN | ago | 0.000 | 2,263 | 826 |
| 10 | TWO | ago | 0.000 | 2,160 | 2,031 |
| 11 | THAT | ago | 0.000 | 1,801 | 722 |
| 12 | IT | ago | 0.000 | 1,695 | 668 |
| 13 | I | ago | 0.000 | 1,694 | 413 |
| 14 | LONG | ago | 0.000 | 1,591 | 1,527 |
| 15 | MONTHS | ago | 0.000 | 1,383 | 1,367 |
| 16 | HE | ago | 0.000 | 1,372 | 240 |
| 17 | HAD | ago | 0.000 | 1,312 | 442 |
| 18 | THREE | ago | 0.000 | 1,187 | 1,110 |
| 19 | SOME | ago | 0.000 | 1,123 | 983 |
| 20 | FEW | ago | 0.000 | 1,084 | 1,039 |
| 21 | YEAR | ago | 0.000 | 1,066 | 980 |

At the bottom of the window, there are tabs for 'concordance', 'collocates', 'plot', 'patterns', 'clusters', 'filenames', and 'source text'. The 'collocates' tab is selected. Below the tabs, it shows '2,871 Type-in AGO'.

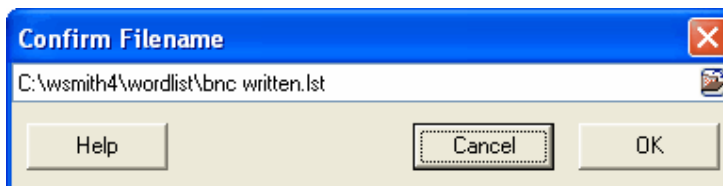
Of the nearly 17,000 instances of *AGO, YEARS* is the top collocate, co-occurring 9,000 times. At this point, only alphabetic sorting or sorting according to frequency is possible.

To measure the strength of each word-pair, the *Mutual Information* score is established:

Choose *Compute | Mutual Information*:



and select a suitable wordlist for the comparison:



Then sort the list by clicking on the *Relation*-column:

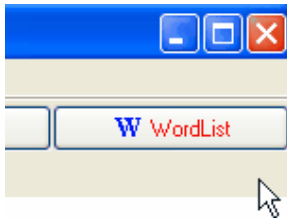
| N | Word | With | Relation | Total | tal Left |
|----|-------------|------|----------|--------|----------|
| 1 | AGO | ago | 12.403 | 16,785 | 47 |
| 2 | HENSLEY | ago | 10.631 | 5 | 1 |
| 3 | AEONS | ago | 9.879 | 11 | 8 |
| 4 | FORTNIGHT | ago | 9.336 | 121 | 121 |
| 5 | YEARS | ago | 9.218 | 9,033 | 8,936 |
| 6 | MOONS | ago | 8.840 | 13 | 12 |
| 7 | WEEKS | ago | 8.754 | 1,047 | 1,029 |
| 8 | SEASONS | ago | 8.548 | 81 | 81 |
| 9 | MILLENNIA | ago | 8.512 | 9 | 9 |
| 10 | MONTHS | ago | 8.387 | 1,383 | 1,367 |
| 11 | MOMENTS | ago | 8.367 | 179 | 178 |
| 12 | UNTHINKABLE | ago | 8.128 | 18 | 15 |
| 13 | DECADE | ago | 7.939 | 165 | 164 |
| 14 | COUPLE | ago | 7.697 | 360 | 342 |
| 15 | TWENTY | ago | 7.658 | 405 | 387 |
| 16 | CENTURIES | ago | 7.592 | 126 | 123 |
| 17 | TEN | ago | 7.521 | 485 | 468 |
| 18 | FIFTY | ago | 7.500 | 133 | 127 |
| 19 | TH | ago | 7.485 | 10 | 0 |
| 20 | MOOTED | ago | 7.471 | 5 | 5 |
| 21 | EIGHTEEN | ago | 7.466 | 54 | 50 |
| 22 | INCEPTION | ago | 7.427 | 9 | 8 |
| 23 | HUNDRED | ago | 7.343 | 250 | 241 |
| 24 | FIFTEEN | ago | 7.342 | 97 | 95 |
| 25 | ANFIELD | ago | 7.340 | 0 | 0 |

The higher the number in the *Relation*-column, the stronger the collocation. The top items in the list now reflect the tendency of *AGO* to co-occur with periods of time and numbers.

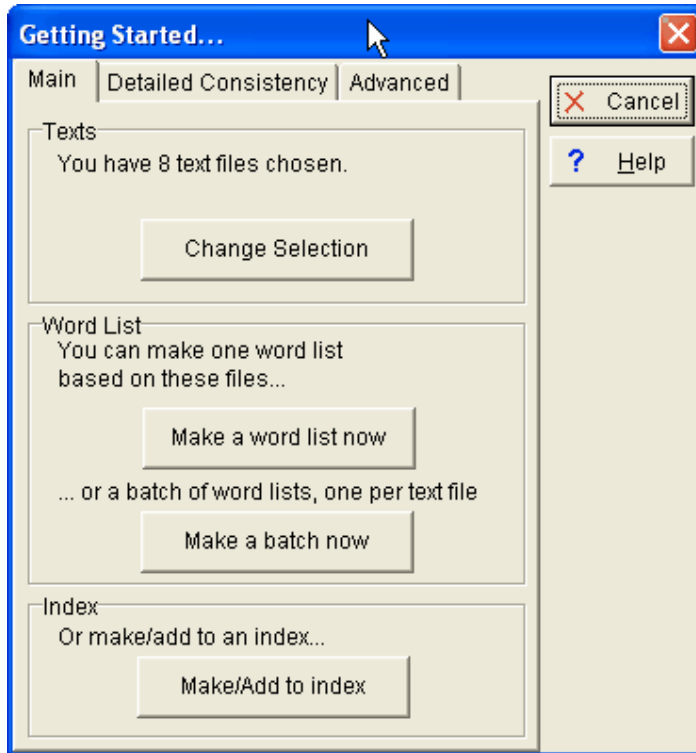
IV. WordList

A *WordList* in WordSmith shows the absolute number of occurrences of each word in the text files, this number converted into percentage of running words, and the number of text files the words occur in.

To create a *WordList*, first click on the button in the main controller:



When *WordList* opens, choose your texts and press *Make a word list now*:



The *WordList* shows a frequency listing ("#" by default is used to represent any number):

The screenshot shows the WordList window with a table of word frequencies. The table has columns for rank (N), word, frequency (Freq.), percentage (%), and number of texts. The most frequent words are listed below:

| N | Word | Freq. | % | Texts |
|---|------|-------|------|-------|
| 1 | # | 985 | 8.90 | 8 |
| 2 | THE | 279 | 2.52 | 8 |
| 3 | AND | 264 | 2.38 | 8 |
| 4 | I | 230 | 2.08 | 7 |
| 5 | TO | 192 | 1.73 | 8 |
| 6 | OF | 187 | 1.69 | 8 |

At the bottom of the window, there are tabs for 'frequency', 'alphabetical', 'statistics', 'filenames', and 'notes'. The 'frequency' tab is currently selected. The status bar shows '2,021' and 'Type-in'.

The most frequent words, besides numbers, are given (the, and, I etc.), the percentage of running words, and how many texts a particular word occurs in.

To have the words alphabetised, click on the *alphabetical*- tab at the bottom of the window.

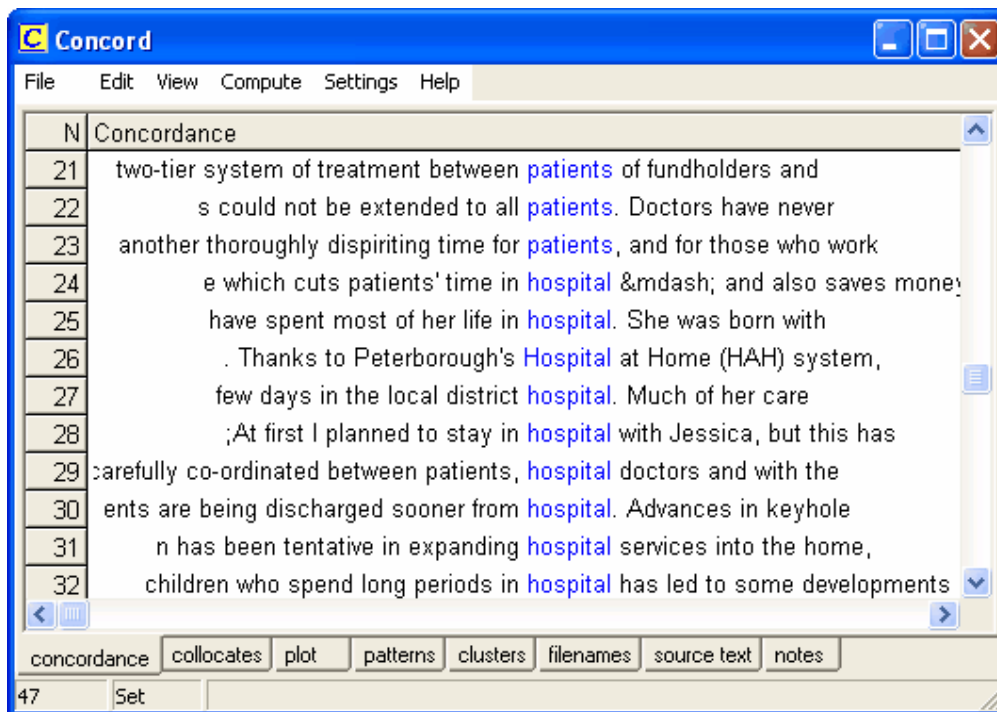
Concordancing Selected Words

Select a word (or more), choose *Compute | Concordance* and you will get the words in their respective contexts:

The screenshot shows the WordList window with a concordance for the word 'PATIENTS'. The 'Compute' menu is open, and 'Concordance' is selected. The table shows the following data:

| N | Word | Freq. | % | Texts |
|----|----------|-------|------|-------|
| 16 | IS | 24 | 0.73 | 1 |
| 17 | AT | 24 | 0.73 | 1 |
| 18 | TH | 24 | 0.73 | 1 |
| 19 | AT | 23 | 0.70 | 1 |
| 20 | ON | 23 | 0.70 | 1 |
| 21 | PATIENTS | 23 | 0.70 | 1 |
| 22 | HOSPITAL | 22 | 0.67 | 1 |
| 23 | HAVE | 19 | 0.58 | 1 |
| 24 | I | 17 | 0.52 | 1 |
| 25 | OUT | 16 | 0.49 | 1 |
| 26 | WAS | 16 | 0.49 | 1 |
| 27 | HOME | 15 | 0.46 | 1 |

The 'Concordance' menu option is highlighted. The status bar shows '1,156' and 'Type-in PATIENTS'.



WordList Statistics

To get statistical results, click on the *statistics*-tab at the bottom of the word list:

| | N | 0 | 1 | 2 | 3 |
|----------------------------------|----------|------------|------------|------------|------------|
| text file | overall | \rom01.txt | \rom02.txt | \rom03.txt | \rom04.txt |
| file size | 56,316 | 4,310 | 12,786 | 5,329 | |
| tokens (running words) in text | 11,073 | 689 | 2,532 | 1,062 | |
| tokens used for word list | 10,088 | 668 | 2,295 | 957 | |
| types (distinct words) | 2,021 | 355 | 687 | 397 | |
| type/token ratio (TTR) | 20.03 | 53.14 | 29.93 | 41.48 | |
| standardised TTR | 37.41 | * | 35.80 | * | |
| standardised TTR std.dev. | 53.90 | * | 45.40 | * | |
| standardised TTR basis | 1,000 | 1,000 | 1,000 | 1,000 | |
| mean word length (in characters) | 4.14 | 5.00 | 4.15 | 4.15 | |
| word length std.dev. | 1.97 | 2.59 | 1.95 | 1.89 | |
| sentences | 171 | 13 | 39 | 13 | |
| mean (in words) | 58.99 | 51.38 | 58.85 | 73.62 | |
| std.dev. | 73.01 | 53.71 | 86.88 | 89.40 | |
| paragraphs | 133 | 3 | 31 | 9 | |
| mean (in words) | 75.85 | 222.67 | 74.03 | 106.33 | |
| std.dev. | 120.58 | 281.95 | 110.33 | 99.61 | |
| headings | | | | | |
| mean (in words) | * | * | * | * | |
| std.dev. | * | * | * | * | |
| sections | 8 | 1 | 1 | 1 | |
| mean (in words) | 1,261.00 | 668.00 | 2,295.00 | 957.00 | |
| std.dev. | 754.96 | * | * | * | |

frequency alphabetical **statistics** filenames notes

2,021 Type-in WHEREFORE

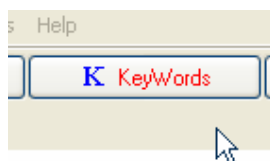
| N | 0 | 1 | 2 |
|-------------------------|-------|-----|-----|
| numbers removed | 985 | 21 | 237 |
| stoplist tokens removed | | | |
| stoplist types removed | | | |
| 1-letter words | 537 | 17 | 103 |
| 2-letter words | 1,507 | 94 | 363 |
| 3-letter words | 1,892 | 126 | 418 |
| 4-letter words | 2,599 | 88 | 575 |
| 5-letter words | 1,680 | 91 | 422 |
| 6-letter words | 721 | 81 | 145 |
| 7-letter words | 478 | 57 | 112 |
| 8-letter words | 355 | 47 | 94 |
| 9-letter words | 169 | 27 | 36 |
| 10-letter words | 83 | 20 | 16 |
| 11-letter words | 24 | 7 | 2 |
| 12-letter words | 16 | 4 | 3 |
| 13-letter words | 13 | 6 | 2 |
| 14-letter words | 7 | 2 | 1 |
| 15-letter words | 5 | 1 | 2 |
| 16-letter words | 1 | | |

frequency alphabetical statistics filenames notes
2,021 Type-in WHEREFORE

V. KeyWords

Keywords are words which occur unusually frequent in a text. This frequency is compared to the keyword's frequency in some kind of reference corpus.

To create a *KeyWords*-list, click on the button in the main controller:

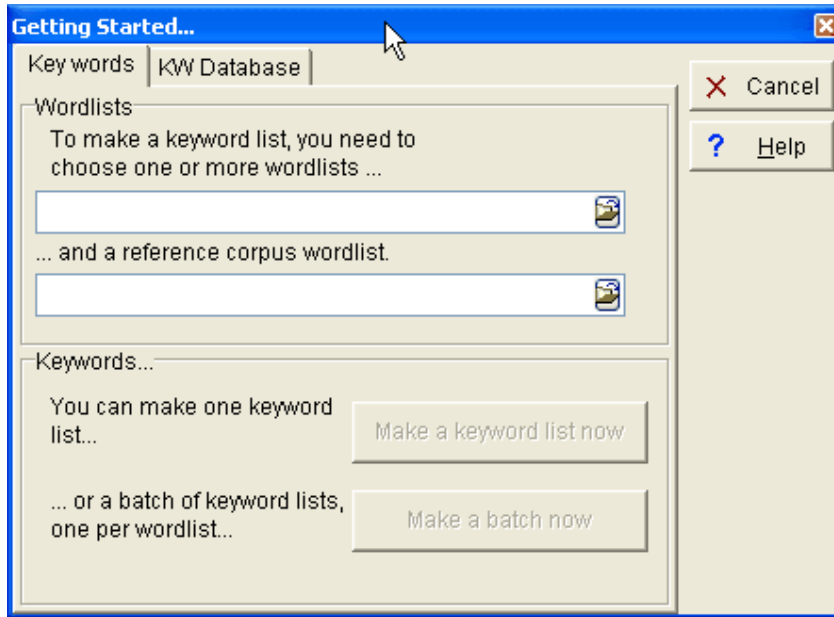


When *KeyWords* opens, choose wordlists by pressing this button:



The reference-corpus wordlist should be big enough to be able to work out significant differences!

Once you have chosen a wordlist and another for your reference, press *Make a keyword list now*:

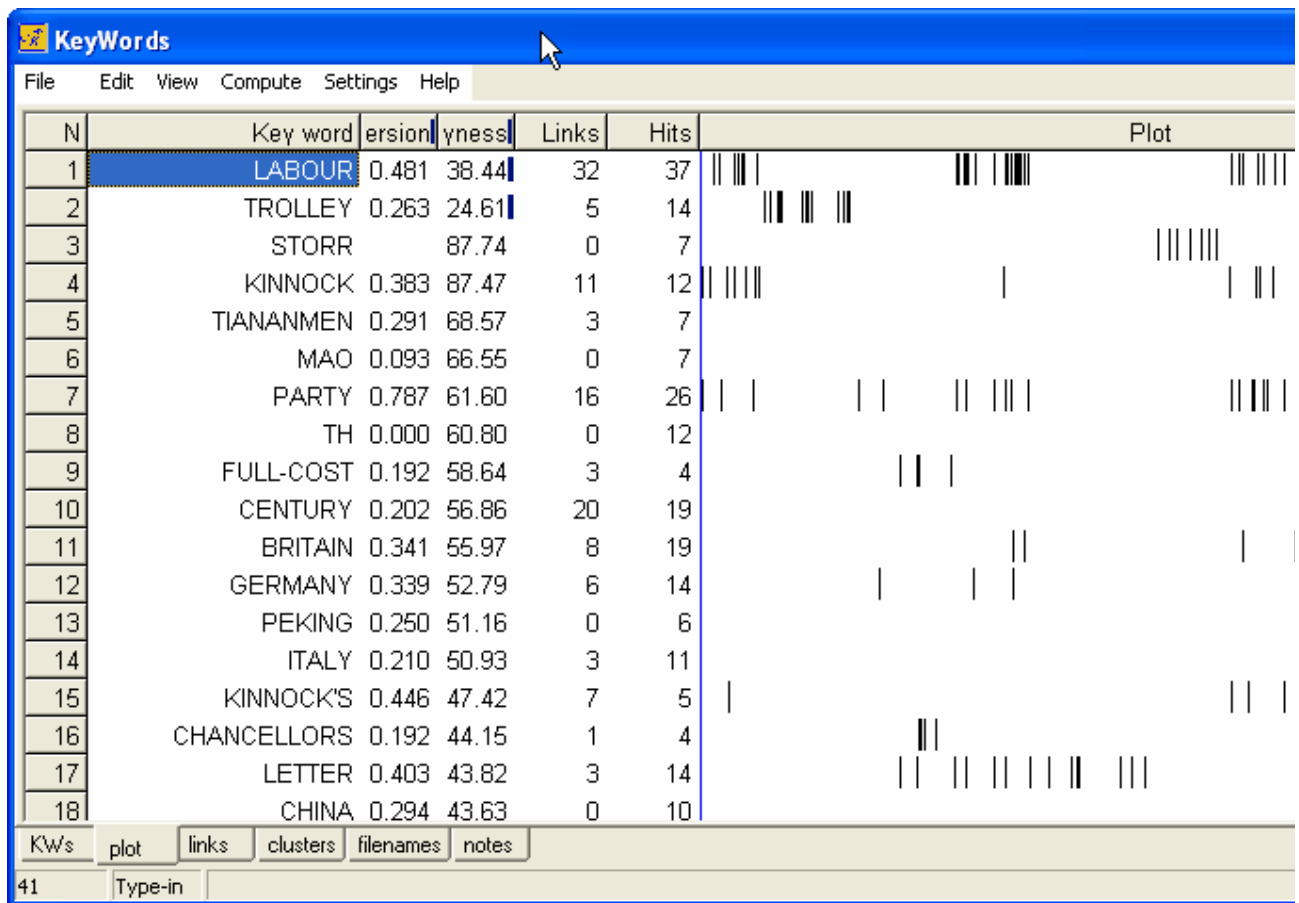


Besides the keyword, its frequency in the source text(s) is given, as well as its frequency in the reference corpus (here, in the BNC).

| N | Key word | Freq. | % | . Freq. | RC. % | eyr |
|----|-----------|-------|------|---------|-------|-----|
| 1 | ROMEO | 130 | 1.17 | 311 | | 83 |
| 2 | BENVOLIO | 49 | 0.44 | 4 | | 86 |
| 3 | THOU | 74 | 0.67 | 753 | | 84 |
| 4 | JULIET | 74 | 0.67 | 1,126 | | 79 |
| 5 | CAPULET | 33 | 0.30 | 14 | | 54 |
| 6 | THEE | 44 | 0.40 | 630 | | 47 |
| 7 | MERCUTIO | 26 | 0.23 | 22 | | 40 |
| 8 | THY | 37 | 0.33 | 632 | | 38 |
| 9 | MONTAGUE | 28 | 0.25 | 134 | | 36 |
| 10 | LOVE | 72 | 0.65 | 22,224 | 0.02 | 34 |
| 11 | TIS | 29 | 0.26 | 423 | | 31 |
| 12 | CAPULET'S | 17 | 0.15 | 0 | | 30 |
| 13 | TYBALT | 17 | 0.15 | 4 | | 28 |
| 14 | SAMSON | 22 | 0.20 | 158 | | 26 |
| 15 | NURSE | 36 | 0.33 | 3,175 | | 26 |

KeyWords Plot

To create a keyword plot, click on the *plot*-tab at the bottom of the window:



The measure of a keyword's dispersion and its keyness are given, as well as the number of links with other keywords of the same text, the absolute number of hits in the text, and a plot-graphic showing the keyword's position in the text.

Concordancing Selected Keywords

Select a keyword (or more), choose *Compute | Concordance* and you will get the keywords in their respective contexts:

| N | | ness | Links | Hits | |
|----|----------------|-------|-------|------|----|
| 1 | | 5.23 | 14 | 24 | |
| 2 | | 9.01 | 14 | 12 | |
| 3 | | 5.08 | 29 | 23 | |
| 4 | | 4.70 | 20 | 22 | |
| 5 | DENPLAN | 0.250 | 03.60 | 4 | 6 |
| 6 | DENTISTS | 0.272 | 91.20 | 2 | 8 |
| 7 | HAH | 0.413 | 81.18 | 11 | 6 |
| 8 | TREATMENT | 0.355 | 65.24 | 5 | 13 |
| 9 | PETERBOROUGH'S | 0.000 | 60.54 | 0 | 4 |
| 10 | NURSE | 0.429 | 53.51 | 11 | 8 |
| 11 | INSURANCE | 0.294 | 48.24 | 6 | 9 |
| 12 | DENTIST | 0.250 | 47.31 | 2 | 5 |
| 13 | DISTRICT | 0.448 | 46.26 | 8 | 9 |
| 14 | HOSPITALS | 0.219 | 46.16 | 5 | 7 |
| 15 | GPS | 0.149 | 44.41 | 2 | 5 |
| 16 | HEALTH | 0.619 | 42.36 | 3 | 12 |
| 17 | DOCTORS | 0.414 | 40.38 | 0 | 7 |

aj5_nhs.kws

File Edit View Compute Settings Help

Concordance

New column
Lemma matches
Matches
Plot

KWs plot links clusters filenames notes

34 Type-in HAH

Concord

File Edit View Compute Settings Help

N Concordance

1 Peterborough's Hospital at Home (HAH) system, Jessica has spent

2 t any one time, Peterborough's HAH has about 24 child or adult

3 ict nurse. Peterborough's HAH was set up more than 10 years

4 Care, singles out Peterborough HAH early discharge scheme for

5 dav. Allen stresses that HAH, which takes about 400 patients

concordance collocates plot patterns clusters filenames source text notes

6 Set