WordSmith Tools Manual

Version 6.0

Lexical Analysis Software Ltd.
Stroud, Gloucestershire, UK

WordSmith Tools Manual

version 6.0

by Mike Scott

2015

WordSmith Tools Manual

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: December 2015

Publisher

Lexical Analysis Software Ltd.

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

Part XII Troubleshooting 454

Part XIII Error Messages 460

WordSmith Tools

Section I

1 WordSmith Tools

WordSmith Tools is an integrated suite of programs for looking at how words behave in texts. You will be able to use the tools to find out how words are used in your own texts, or those of others.

The **WordList** tool lets you see a list of all the words or word-clusters in a text, set out in alphabetical or frequency order. The concordancer, **Concord**, gives you a chance to see any word or phrase in context -- so that you can see what sort of company it keeps. With **KeyWords** you can find the key words in a text.

The tools have been used by Oxford University Press for their own lexicographic work in preparing dictionaries, by language teachers and students, and by researchers investigating language patterns in lots of different languages in many countries world-wide.

Getting Help

Online step-by-step screenshots showing what WordSmith does.

Most of the menus and dialogue boxes have help options. You can often get help just by pressing F1 or ? , or by choosing Help (at the right hand side of most menus). Within a help file (like this one) you may find it easiest to click the Search button and examine the index offered, or else just browse through the help screens.

See also: getting started straight away with WordList 17, Concord 14, or KeyWords 15.

Version:6.0

Overview

Section II

2 Overview

2.1 Requirements

WordSmith Tools requires

1. a reasonably up-to-date computer 440
2. running Windows XP 440 or later
3. your own collection of text in plain text format or converted 365 to plain text

2.2 What's new in version 6

WordSmith is organic software!

Version 5.0 was started in June 2007, three years after version 4.0 and has continued this organic policy of growth ever since ... now in 2015 we are at version 6.0 with improvements and new features.

New features:

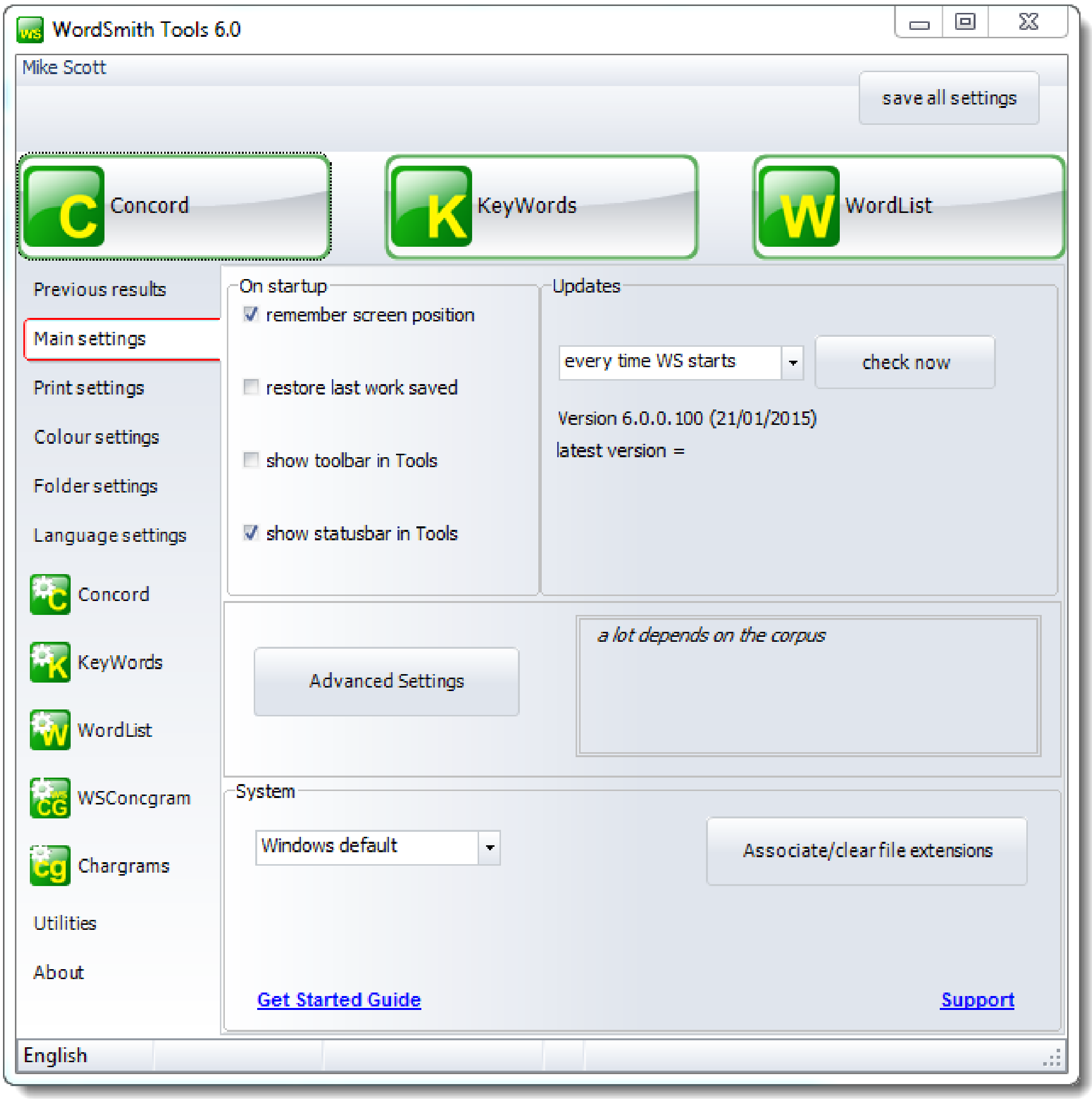
- Move files to sub-folders 351
- Skins 60
- Word Clouds 128
- Date handling & Time-lines 126
- .docx files 365
- Scripting 106
- Colour categories 51
- Phrase frames 281
- Collocate following 184
- Chargrams 409

2.3 Controller

This program controls the Tools. It is the one which shows and alters current defaults, handles the choosing of text files, and calls up the different Tools.

It will appear at the top left corner of your screen.

You can minimise it, if you feel the screen is getting cluttered 127.



WordSmith Tools 6.0

Mike Scott

On startup

☑ remember screen position

□ restore last work saved

□ show toolbar in Tools

☑ show statusbar in Tools

Updates

every time WS starts

Version 6.0.0.100 (21/01/2015)

latest version =

a lot depends on the corpus

System

Windows default

English

For a step-by-step view with screenshots, click here to visit the WordSmith website.

2.4 Concord

Concord is a program which makes a concordance 159 using plain text or web text files.

To use it you will specify a search word, which Concord will seek in all the text files you have chosen. It will then present a concordance display, and give you access to information about collocates of the search word.

Listings can be saved 211 for later use, edited, printed, copied to your word-processor, or saved as text files.

See also: Concord Help Contents Page 158, The buttons 441

2.5 KeyWords

The purpose of this program is to locate and identify key words in a given text. To do so, it compares the words in the text with a reference set of words usually taken from a large corpus of text. Any word which is found to be outstanding in its frequency in the text is considered "key". The key words are presented in order of outstandingness.

The distribution of the key words can be plotted 251.

Listings can be saved 101 for later use, edited, printed, copied to your word-processor, or saved as text files.

This program needs access to 2 or more word lists, which must be created first, using the Word List 6 program.

See also: KeyWords Help Contents Page 229, The buttons 441

2.6 WordList

WordList generates word lists based on one or more plain text 419 or web text files. Word lists are shown both in alphabetical and frequency order. They can be saved 101 for later use, edited, printed, copied to your word-processor, or saved as text files.

See also: WordList Help Contents Page 258

2.7 Utilities

2.7.1 Character Profiler

A tool to help find out which characters or chargrams 426 are most frequent in a text or a set of texts.

The purpose could be to check out which characters or character sequences are most frequent (e.g. in normal English text the letter **E** followed by **T** will be most frequent, **THE** and **ARE** will be high frequency 3-chargrams), or it could be to check whether your text collection contains any oddities, such as accented characters or curly apostrophes you weren't expecting.

See also: Character Profiling 405

2.7.2 CharGrams

A tool to help find out which chargrams 426 (sequences of characters) are most frequent in a text or a set of texts.

The purpose could be to check out which chargrams are most frequent e.g. in word-initial position, in the middle of a word, or at the end.

See also: Chargrams Tool 409

2.7.3 Choose Languages

A tool for selecting Languages which you want to process.
You will probably only need to do this once, when you first use WordSmith Tools.

See also: Choose Language Tool 84

2.7.4 Corpus Corruption Detector

A tool to go through your corpus and seek out any text files which may have become corrupted. Works in any language.

See also: detecting corpus corruption 329

2.7.5 File Utilities

Utilities to

- compare two files 347
- cut large files into chunks 348
- find duplicate files 348
- rename 349 multiple files
- split large files into their component texts 343
- join up 346 a lot of small text files into merged text files
- find holes 354 in your text files

2.7.5.1 Splitter

Splitter is a component of File Utilities which splits large files into small ones for text analysis purposes. You can specify a symbol to represent the end of a text (e.g. **`</Text>`**) and Splitter will go through a large file copying the text; each time it finds the symbol it will start a new text file.

See also: Splitter Help Contents Page 343

2.7.6 File Viewer

A tool for viewing how your text files are formatted in great detail, character by character.

See also: File Viewer Index 84

2.7.7 Minimal Pairs

a program to find typos and minimally-differing pairs of words.

See also : aim 331, requirements 332, choosing your files 333, output 336, rules and settings 337, running the program 338.

2.7.8 Text Converter

Text Converter is a general-purpose utility which you use for three main tasks: to edit your texts, to rename text files, to change file attributes, to move files into a new folder if they contain certain words or phrases.

The main use is to replace strings in text files. It does a "search and replace" much as in word-processors, but it can do this on lots of text files, one after the other. As it does so, it can also replace any number of strings, not just one.

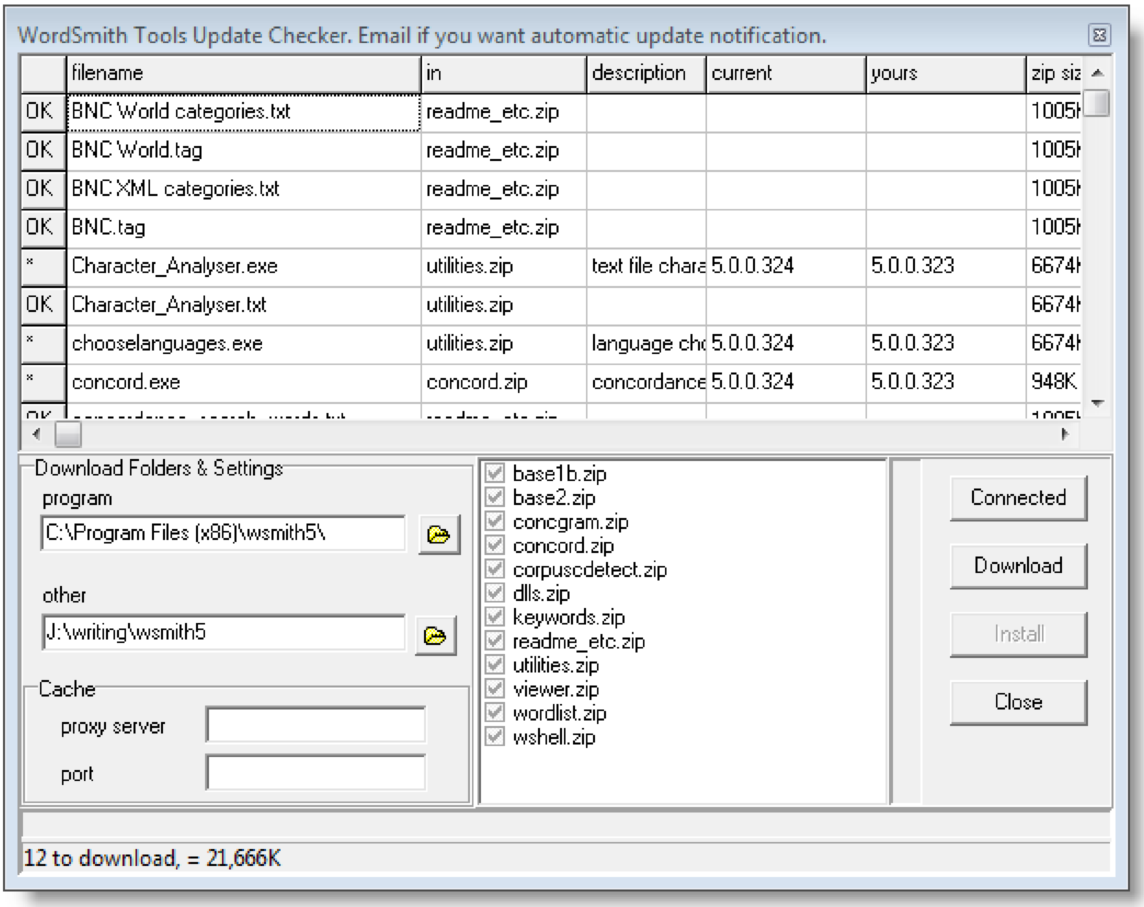
It is very useful for going through large numbers of texts and re-formatting them as you prefer, e.g. taking out unnecessary spaces, ensuring only paragraphs have <Enter> at their ends, changing accented characters.

See also: Text Converter Help Contents Page 355

2.7.9 Version Checker

A tool to check whether any components of your current version need updating and if so, download them for you. Accessed via the main Controller menu, *File* | *Web version check*.

When you run the program you'll see something like this:



WordSmith Tools Update Checker. Email if you want automatic update notification.

	filename	in	description	current	yours	zip siz
OK	BNC World categories.txt	readme_etc.zip				1005K
OK	BNC World.tag	readme_etc.zip				1005K
OK	BNC XML categories.txt	readme_etc.zip				1005K
OK	BNC.tag	readme_etc.zip				1005K
*	Character_Analyser.exe	utilities.zip	text file chara	5.0.0.324	5.0.0.323	6674K
OK	Character_Analyser.txt	utilities.zip				6674K
*	chooselanguages.exe	utilities.zip	language ch	5.0.0.324	5.0.0.323	6674K
*	concord.exe	concord.zip	concordance	5.0.0.324	5.0.0.323	948K

Download Folders & Settings

program: C:\Program Files (x86)\wsmith5\

other: J:\writing\wsmith5

Cache

proxy server:

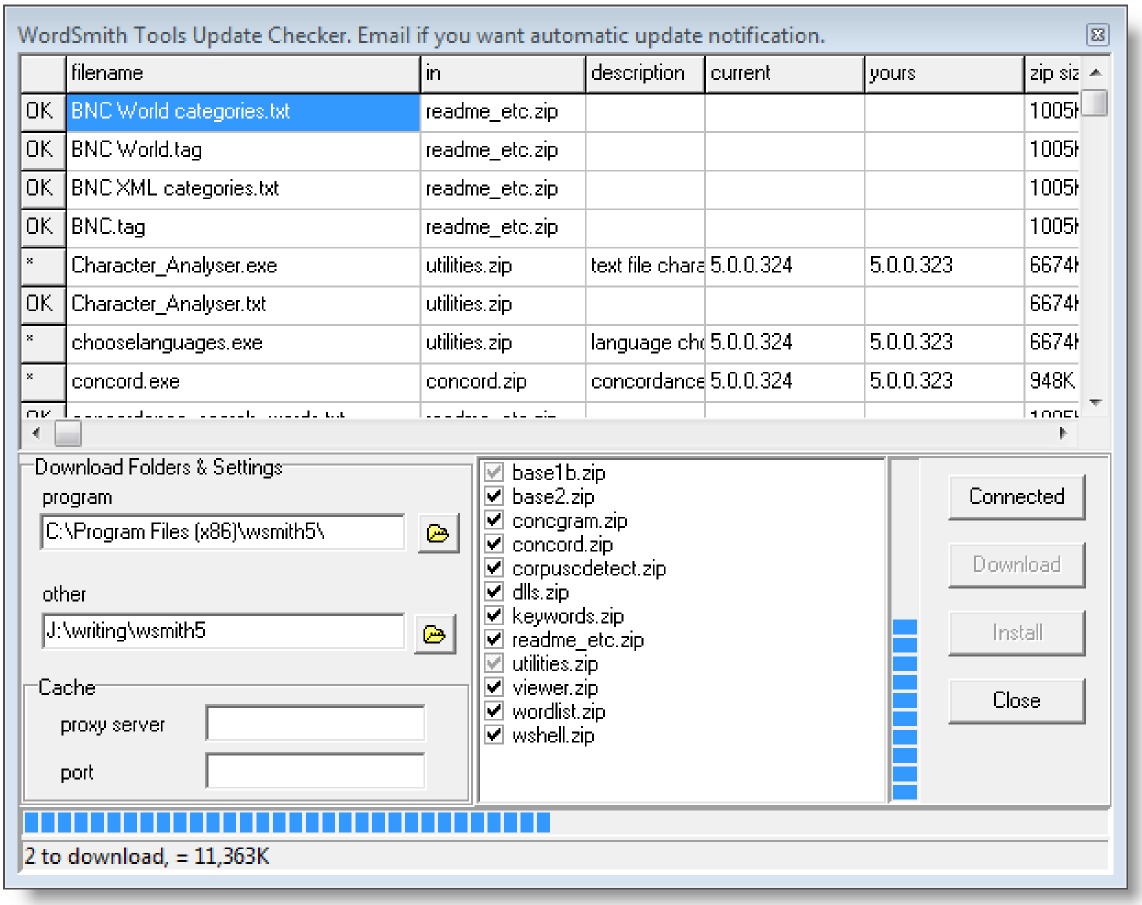
port:

☑ base1b.zip
☑ base2.zip
☑ concgram.zip
☑ concord.zip
☑ corpuscdetect.zip
☑ dlls.zip
☑ keywords.zip
☑ readme_etc.zip
☑ utilities.zip
☑ viewer.zip
☑ wordlist.zip
☑ wshell.zip

12 to download, = 21,666K

The various components of WordSmith are listed in the top window and the current version is compared with your present situation. If they are different, all the files in the relevant zip file will be starred (*) in the left margin.

By default you will download to wherever WordSmith is already (the program in a program folder and settings etc. in a Documents folder) but you're free to choose somewhere else. Press Download if you wish to get the updated files.



WordSmith Tools Update Checker. Email if you want automatic update notification.

	filename	in	description	current	yours	zip siz
OK	BNC World categories.txt	readme_etc.zip				1005K
OK	BNC World.tag	readme_etc.zip				1005K
OK	BNC XML categories.txt	readme_etc.zip				1005K
OK	BNC.tag	readme_etc.zip				1005K
*	Character_Analyser.exe	utilities.zip	text file chara	5.0.0.324	5.0.0.323	6674K
OK	Character_Analyser.txt	utilities.zip				6674K
*	chooselanguages.exe	utilities.zip	language chc	5.0.0.324	5.0.0.323	6674K
*	concord.exe	concord.zip	concordance	5.0.0.324	5.0.0.323	948K

Download Folders & Settings

program: C:\Program Files (x86)\wsmith5\

other: J:\writing\wsmith5

Cache

proxy server:

port:

☑ base1b.zip
☑ base2.zip
☑ concgram.zip
☑ concord.zip
☑ corpuscdetect.zip
☑ dlls.zip
☑ keywords.zip
☑ readme_etc.zip
☑ utilities.zip
☑ viewer.zip
☑ wordlist.zip
☑ wshell.zip

2 to download, = 11,363K

After the download, the various .zip files are checked (bottom right window) if downloaded successfully, and the Install button is now available for use. Install unzips all those which are checked.

2.7.10 Viewer and Aligner

Viewer & Aligner is a utility which enables you to examine your files in various formats. It is called on by other Tools whenever you wish to see the source text.

Viewer & Aligner can also be used simply to produce a copy of a text file with numbered sentences or paragraphs 386 or for aligning 381 two or more versions of a text, showing alternate paragraphs or sentences of each.

See also: Viewer & Aligner Help Contents Page 380

2.7.11 Webgetter

A tool to gather text from the Internet.

The point of it…
The idea is to build up your own corpus of texts, by downloading web pages with the help of a search engine.

See also: A fuller overview 324, Settings 325, Display 326, Limitations 328

2.7.12 WSConcGram

a tool for generating concgrams 394.

See also : Aims of WSConcGram 393, Running WSConcGram 395

Getting Started

Section III

3 Getting Started

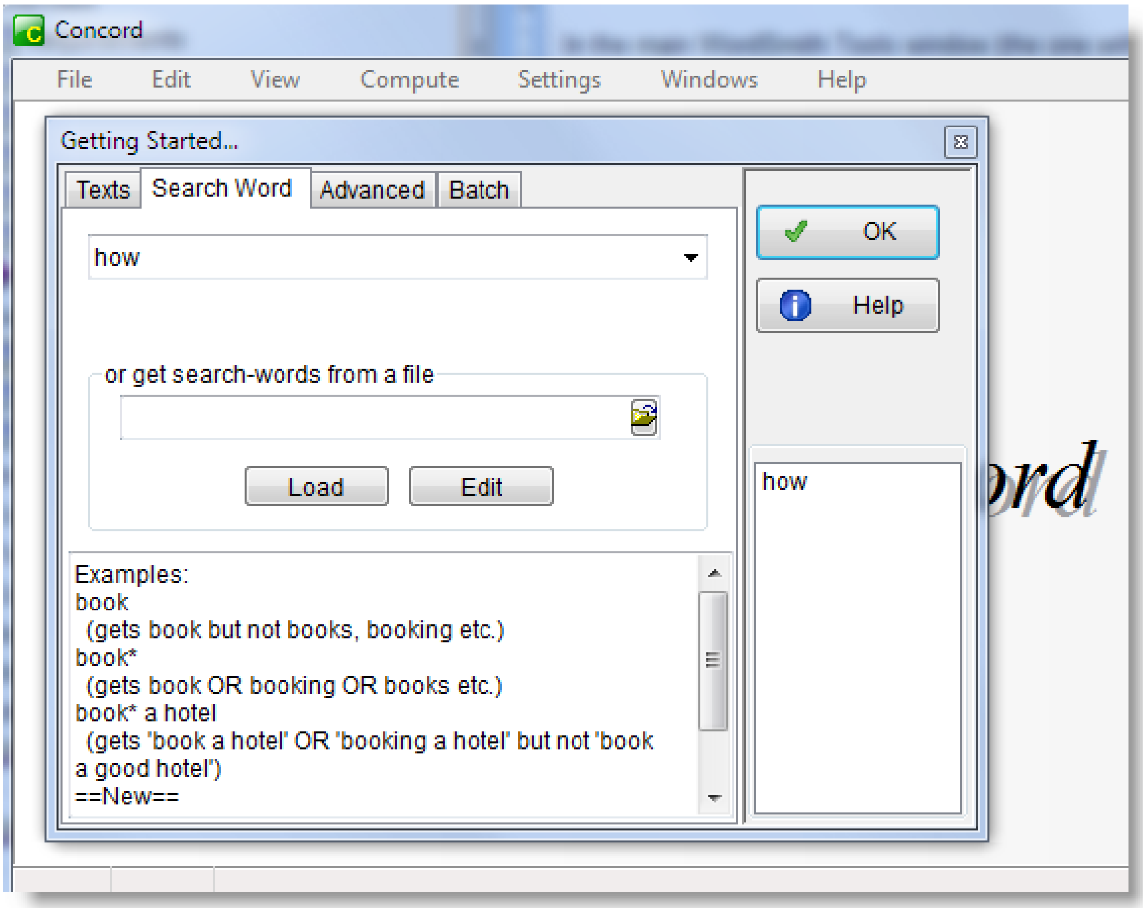
3.1 getting started with Concord

For a step-by-step view with screenshots, visit the WordSmith website.

In the main WordSmith Tools window (the one with WordSmith Tools Controller 4 in its title bar), choose the Tools option, and once that's opened up, you'll see the Concord button. Click and the Concord tool will start up.

Choose File | New ●

You should now see a dialogue box which lets you choose your texts 44 or change your choice, and make a new concordance, looking somewhat like this:



how

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==

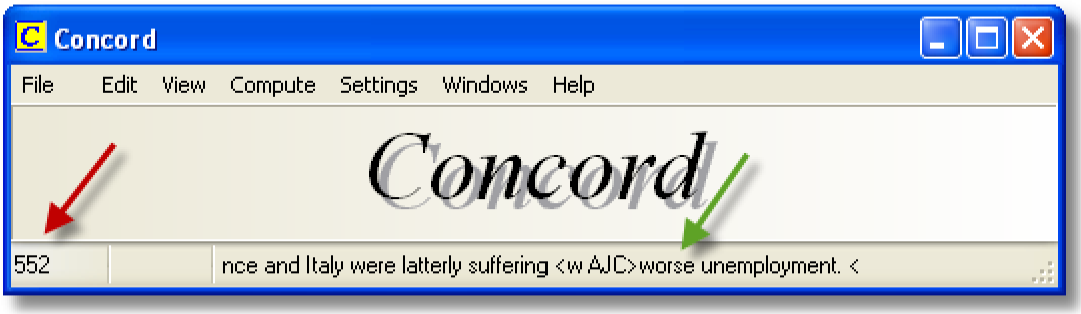
how

(If you only see the window with Concord in its caption, choose *File* | *New* (●) and the Getting Started window will open up.)

If you have never used WordSmith 446 before you will find a text has been selected for you automatically to help you get started.

You will need to specify a Search-Word or phrase 159 and then press OK (✓).

While Concord is working, you may see a progress indicator like this.


Concord
File Edit View Compute Settings Windows Help
Concord
552
nce and Italy were latterly suffering <w AJC>worse unemployment. <

Here, we have 552 entries so far, and the last one in shows the context for *worse*, our search-word.

If you want to alter other settings, press Advanced 222, but you can probably leave the default settings as they are.

Concord now searches through your text(s) looking for the search word or Tag 198.

Don't forget to save the results 211 (press Ctrl+F2 or 💾) if you want to keep the concordance for another time.

See also: Concord Help Contents 158.

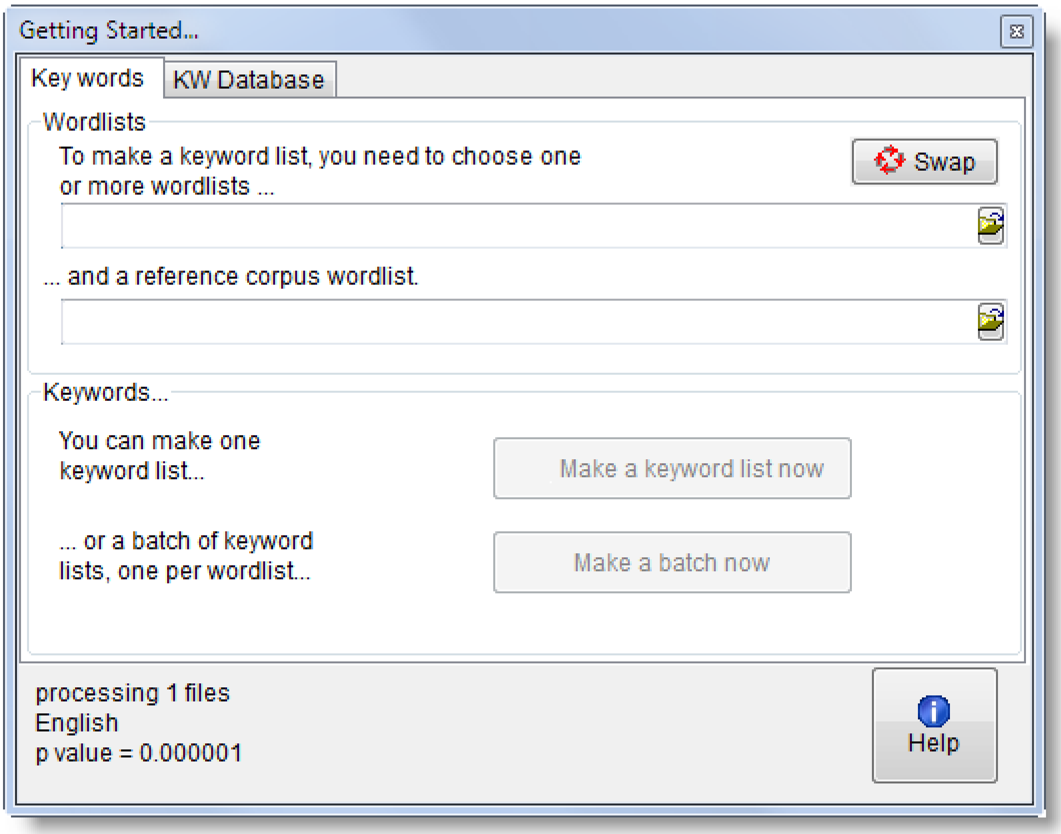
3.2 getting started with KeyWords

For a step-by-step view with screenshots, visit the WordSmith website.

In the main WordSmith Tools window (the one with WordSmith Tools Controller 4 in its title bar), choose the Tools option, and once that's opened up, you'll see KeyWords. Click and KeyWords will open up.

Choose File | New ●

You see a dialogue box which lets you choose your word-lists 232.



Getting Started...

Key words | KW Database

Wordlists

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

... and a reference corpus wordlist.

Keywords...

You can make one keyword list...

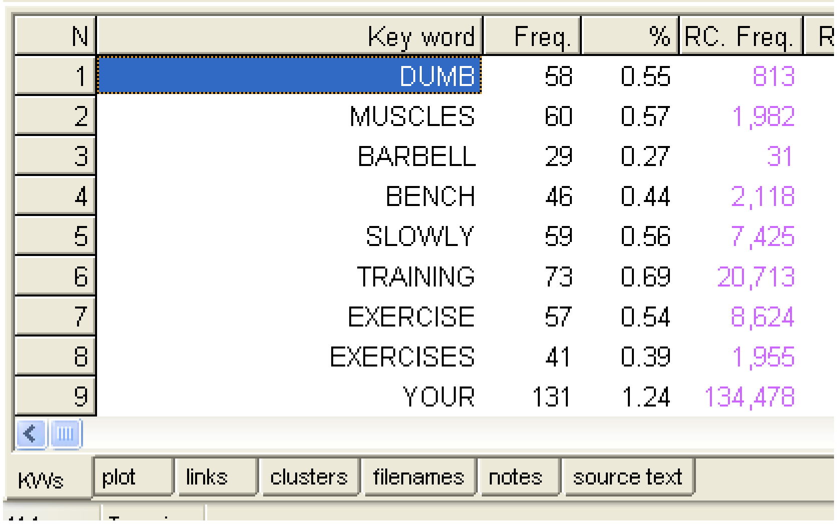
... or a batch of keyword lists, one per wordlist...

processing 1 files
English
p value = 0.000001

You'll need to choose two word lists to make a key words list from: one based on a single text (or single corpus), and another one based on a corpus of texts, enough to make up a good reference corpus for comparison.

You will see two lists of the word list files in your current word-list folder. If there aren't any there, go back to the WordList tool and make some word lists. Choose one small word list above, and a reference corpus 447 list below to compare it with. With your texts selected, you're ready to do a key words analysis. Click on make a keyword list now.

You'll find that KeyWords starts processing your file and a progress 123 window in the main Controller shows a bar indicating how it's getting on. After KeyWords has finished, it will show you a list of the key words. The ones at the top are more "key" than those further down.



N	Key word	Freq.	%	RC. Freq.	R
1	DUMB	58	0.55	813	
2	MUSCLES	60	0.57	1,982	
3	BARBELL	29	0.27	31	
4	BENCH	46	0.44	2,118	
5	SLOWLY	59	0.56	7,425	
6	TRAINING	73	0.69	20,713	
7	EXERCISE	57	0.54	8,624	
8	EXERCISES	41	0.39	1,955	
9	YOUR	131	1.24	134,478	

Don't forget to save the results 101 (press Ctrl+F2) if you want to keep the keyword list for another time.

See also: KeyWords Help Contents 229, What's it for? 229

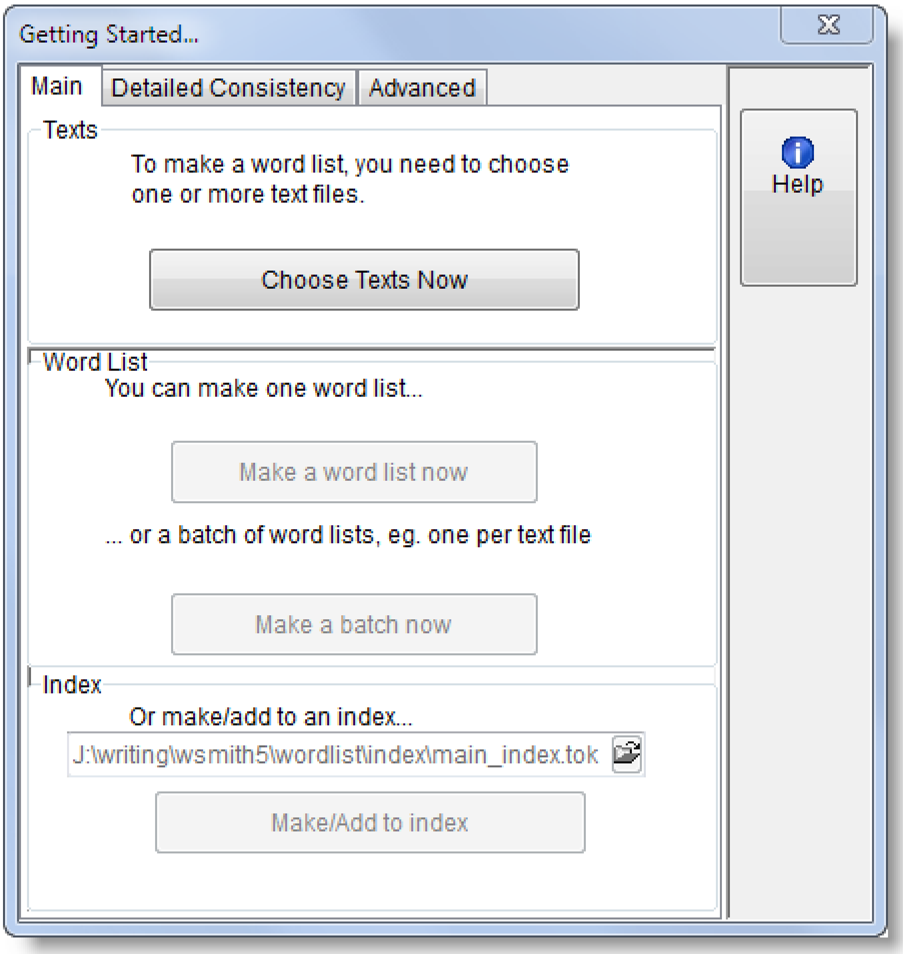
3.3 getting started with WordList

For a step-by-step view with screenshots, visit the WordSmith website.

I suggest you start by trying the WordList program. In the main WordSmith Tools window (the one with WordSmith Tools Controller 4 in its title bar), choose the Tools option, and once that's opened up, you'll see WordList. Click and WordList will open up.

Choose File | New ●

You will see a dialogue box which lets you choose your texts 44 or change your choice, and make a new word list.



Texts

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

Word List

You can make one word list...

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

Index

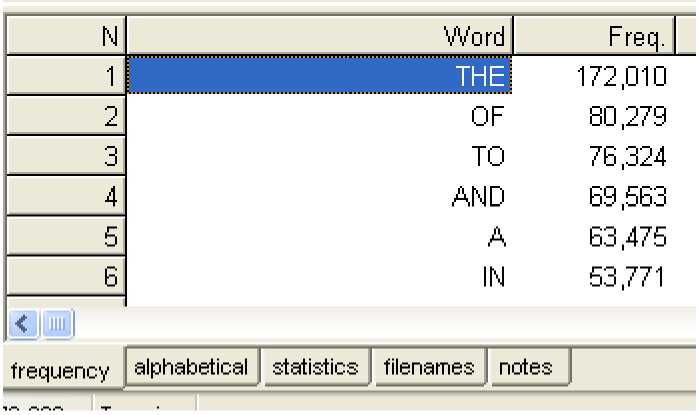
Or make/add to an index...

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

If you have never used WordSmith 446 before you will find a text has been selected for you automatically to help you get started.

There are other settings which can be altered via the menu, but usually you can just go straight ahead and make a new word list, individually or as a Batch 39.

You'll find that WordList starts processing your file(s) and a progress 123 window in the main Controller shows a bar indicating how it's getting on. After WordList has finished making the list, you will see some windows showing the words from your text file in alphabetical order and in frequency order, statistics, filenames, notes 29.



N	Word	Freq.
1	THE	172,010
2	OF	80,279
3	TO	76,324
4	AND	69,563
5	A	63,475
6	IN	53,771

Don't forget to save the results 101 (press Ctrl+F2 or 💾) if you want to keep the word list for another time.

See also: WordList Help Contents 258.

Installation and Updating

Section IV

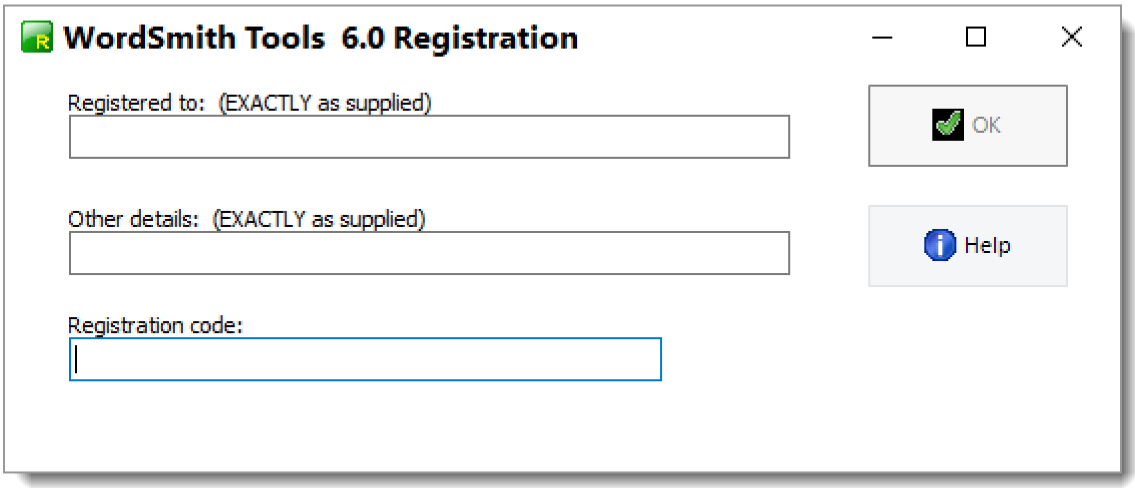
4 Installation and Updating

4.1 installing WordSmith Tools

1. You have run or downloaded and then run one or more .exe files.
2. This will expand all the files needed for WordSmith Tools into the folder of your choice (**`\wsmith6`** by default). You can install to a removable drive if you wish (explained below).
3. Now run **`\wsmith6\wordsmith6.exe`** to get started. You will be asked to register. Otherwise WordSmith will go through its paces as a Demonstration Version 427.

Upon receipt of the registration code, run WordSmith Tools. If you have only just installed the registration program will start up automatically. If not you can run **`\wsmith6\WSRegister6.exe`**.

Single User Licence


WordSmith Tools 6.0 Registration
Registered to: (EXACTLY as supplied)
Other details: (EXACTLY as supplied)
Registration code:
OK
Help

Everything must correspond exactly to what you were given when you purchased.

Paste in your Name as specified in your purchase email or screen and (if there are any in the registration) Other Details, and paste in the code.

This name appears in the main window and whenever you access the About option (F9). Your software will then be fully enabled, and the Update from Demo menu option will disappear. (The **`WSRegister6.exe`** program will still be there in your \wsmith6 folder, and can be used if you ever need to re-register.)

You may require Administrator rights to register an installation using **`\Program Files`**. See this link or search "run as Administrator".

install to a removable drive

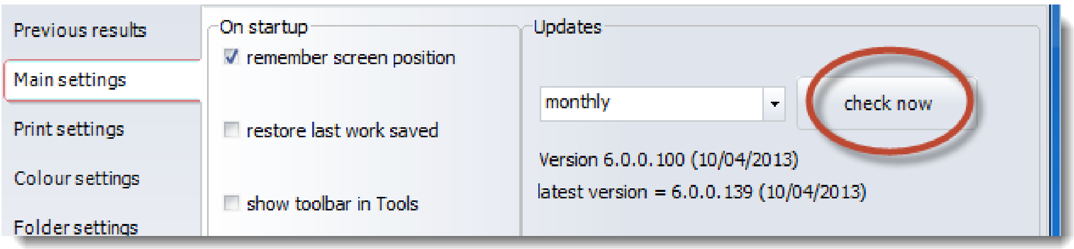
You don't need to install to the **`C:\`** drive -- you can install WordSmith on a USB drive such as a pen drive or memory stick, or a fast external hard drive. That way you can take WordSmith with you from one computer to the next. A pen drive will be a rather slow medium, but a fast external drive can be very satisfactory in terms of speed. If you save your default settings 113, any folder names 431 which are on the external drive itself get the drive letter corrected automatically.

Site Licence

Follow the instructions at http://lexically.net/wordsmith/version6/faqs/network_installation.htm.

Updating your version

To update a version so as to get the very latest build of the program, just check the button in the Updates box.



On startup

☑ remember screen position

□ restore last work saved

□ show toolbar in Tools

Updates

monthly

Version 6.0.0.100 (10/04/2013)

latest version = 6.0.0.139 (10/04/2013)

Or simply re-install afresh with a complete new download.

To update a demo version, visit http://lexically.net/wordsmith/purchasing.htm for details of suppliers.

If you make a mistake and your registration fails, you can try again. You can get a more recent version at the WordSmith home page.

To un-install, just delete all the files in your `\wsmith6` folder. Your data may be in sub-folders of `\wsmith6` or in sub-folders of your `Documents\wsmith6.`

See also: Setting default options 113, Contact Addresses 425, File types 428.

4.2 what your licence allows

In among the legal stuff you will find this, in relation to single user licences:

SINGLE USER LICENCES

Think of these as a licence for a person.

You can install the product on a machine at your office and a machine at home. You may yourself use both copies of the product, but only one at a time.

You cannot install the product on two machines, and then use both of those copies at the same time, or allow anyone else to use your copy of the product on the second machine. For instance, you cannot purchase one copy of the product and allow a friend or family member to use the product on the other machine.

You may not, at any time, allow another user to install your copy of the product for his/her own use.

SITE LICENCES

Think of these as a licence for a given number of terminals.

The full licence text is at **`\wsmith6\user_licence.txt`**. Your licence doesn't expire but it is limited to the major version you bought. So a WordSmith 5 licence won't work with WordSmith 6 and a WS 6 licence won't work with WS7 but you can re-download the program as needed from our site.

4.3 site licence defaults

If you have bought a site licence, just install one copy of WordSmith on any shared drive accessible by all your users. Follow the instructions at http://lexically.net/wordsmith/version6/faqs/network_installation.htm.

the wordsmith6.ini file

This file is in the folder where you installed: in it you will see a section which allows you to specify exactly where each user should save their preferences.

The following terms are used

```
prohibited drives
limited folder
instructions folder
network-read/write folder
```

and an example would be

```
[NETWORK]
network-read/write folder=m:\Documents\wsmith6
```

(drive M: is to be used when running on the network as it's one any user can write to.)

```
prohibited drives=xyz
```

(X: Y: and Z: are drives you don't want your users to look in when choosing texts.)

```
limited folder=v:\texts
```

(V:\TEXTS -- and any sub-directories of it -- is where users will by default choose their corpus on your network; though they may of course look elsewhere in any other drives they control.)

```
instructions folder=L:\English\WSmith instructions
```

(when you run the software in a teaching session, you will put the instructions in that folder.)

When a new user starts using WordSmith for the very first time, WordSmith will notice

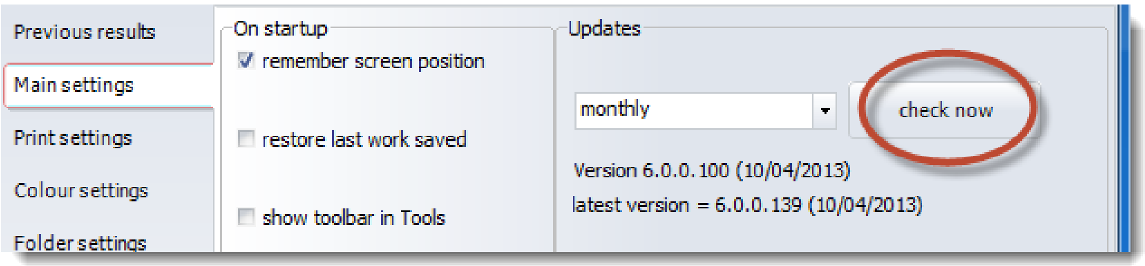
> ***that it is running on a site-licence version and read the "network-read/write folder" information above. It will then try to automatically create the folder you have specified above (in theory you shouldn't need to do it yourself) and copy the various .ini and other settings files over from the folder on your server where the WordSmith program is, to that folder. Your life as an installer will be a lot easier if the drive and folder you specify is truly one your users can write to!***

For networked drives, because of a Microsoft security update involving HTML Help files, WordSmith will copy the **`wordsmith6.chm`** file to the user's Windows-allocated temporary folder.

See also: What your licence allows 22, Class Instructions 51

4.4 version checking

You can check whether your version is up to date in WordSmith's main settings:

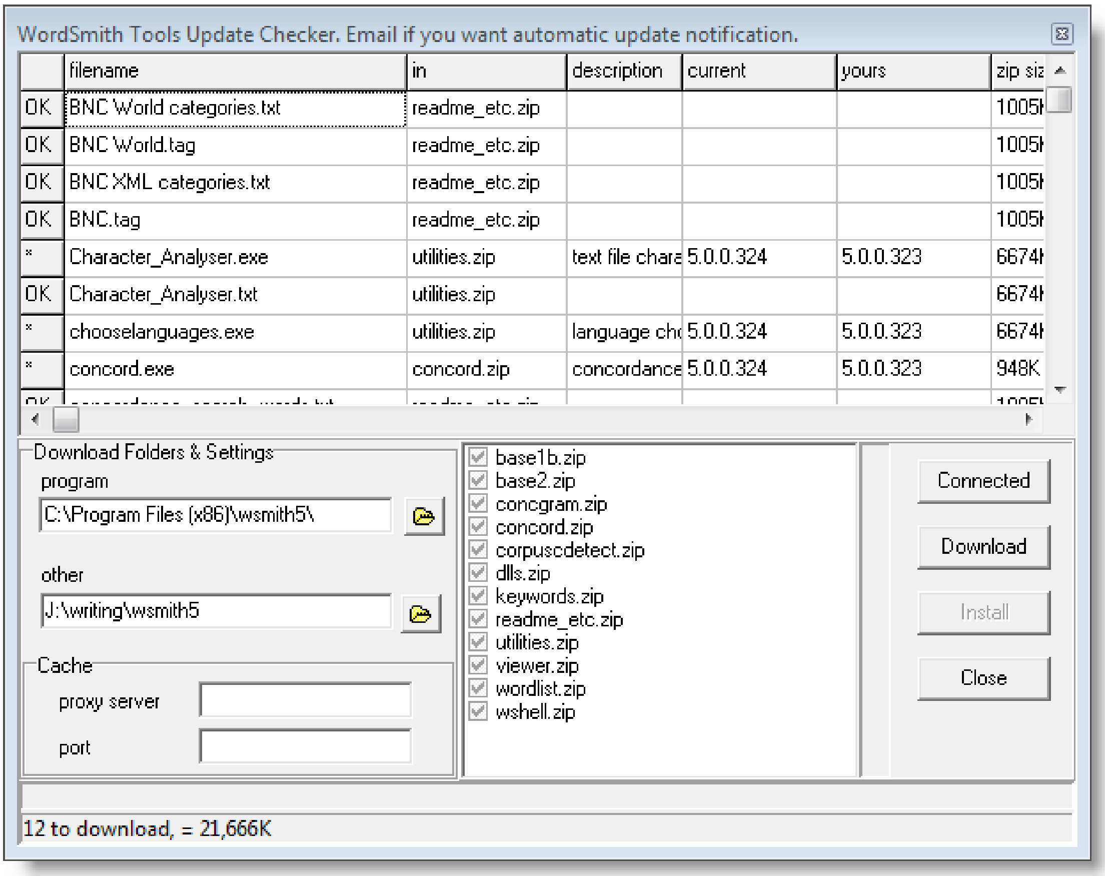


On startup
☑ remember screen position
□ restore last work saved
□ show toolbar in Tools

Updates
monthly
Version 6.0.0.100 (10/04/2013)
latest version = 6.0.0.139 (10/04/2013)

and can set the check to occur regularly every month, week etc.

Besides this, WordSmith comes with a utility (**`wordsmith_version_check.exe)`** which enables you download the necessary upgrades and patches. In order to install these, WordSmith itself will need to close down.



WordSmith Tools Update Checker. Email if you want automatic update notification.

	filename	in	description	current	yours	zip siz
OK	BNC World categories.txt	readme_etc.zip				1005K
OK	BNC World.tag	readme_etc.zip				1005K
OK	BNC XML categories.txt	readme_etc.zip				1005K
OK	BNC.tag	readme_etc.zip				1005K
*	Character_Analyser.exe	utilities.zip	text file chara	5.0.0.324	5.0.0.323	6674K
OK	Character_Analyser.txt	utilities.zip				6674K
*	chooselanguages.exe	utilities.zip	language ch	5.0.0.324	5.0.0.323	6674K
*	concord.exe	concord.zip	concordance	5.0.0.324	5.0.0.323	948K

Download Folders & Settings

program: C:\Program Files (x86)\wsmith5\

other: J:\writing\wsmith5

Cache

proxy server:

port:

☑ base1b.zip
☑ base2.zip
☑ concgram.zip
☑ concord.zip
☑ corpuscdetect.zip
☑ dlls.zip
☑ keywords.zip
☑ readme_etc.zip
☑ utilities.zip
☑ viewer.zip
☑ wordlist.zip
☑ wshell.zip

12 to download, = 21,666K

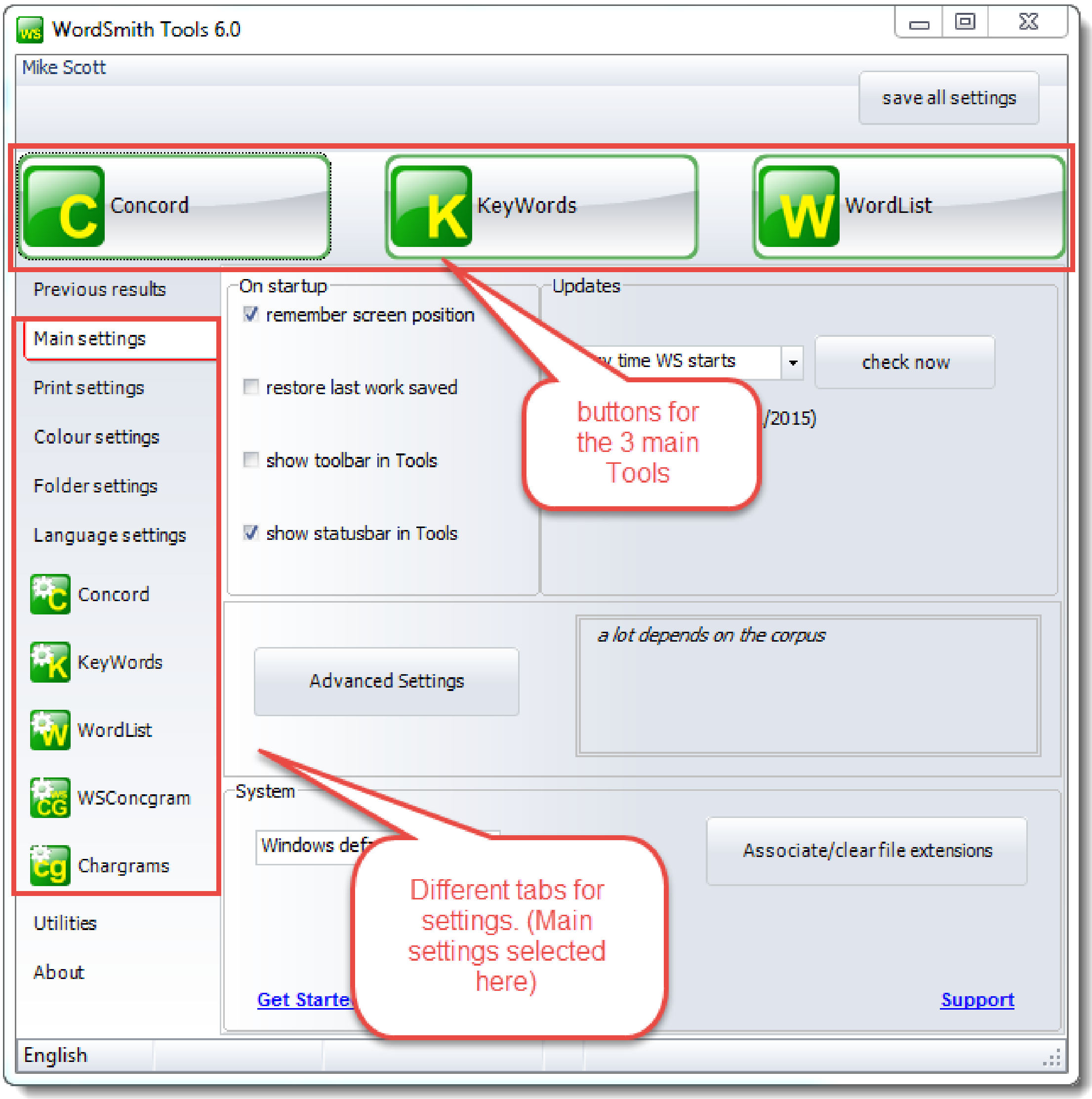
See also: version information 451, version updating 9.

Controller

Section V

5 Controller

The main WordSmith Controller is a window which holds all the numerous settings and behind the scenes tells each Tool what to do. You can start up only one Controller -- though you can start up numerous Concord windows and WordList windows etc.

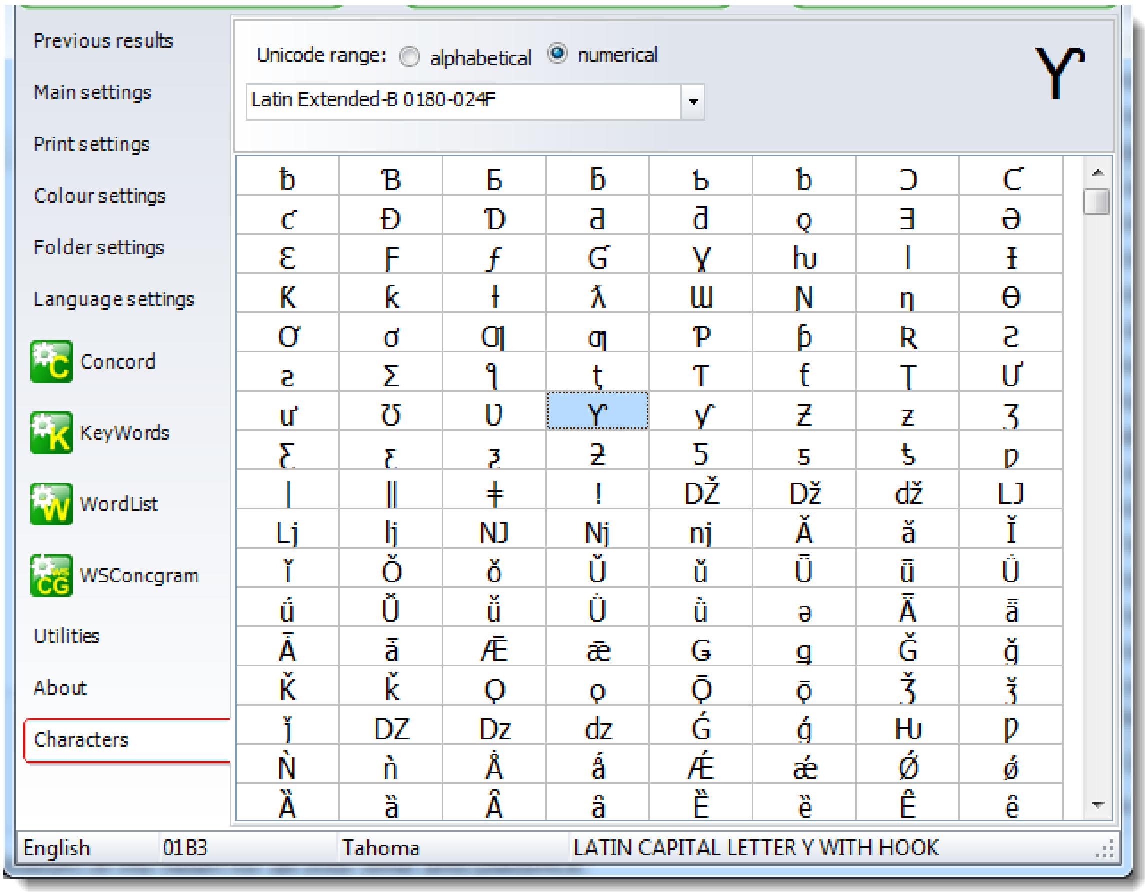


It is best to leave the Controller in one default position on your screen -- there is no advantage in maximizing its size.

5.1 characters and letters

5.1.1 accents and other characters

This window shows a set of the characters available using Unicode.



Unicode range: ○ alphabetical ● numerical

Latin Extended-B 0180-024F

ƀ	Ɓ	Ƃ	ƃ	Ƅ	ƅ	Ɔ	Ƈ
ƈ	Ɖ	Ɗ	Ƌ	ƌ	ƍ	Ǝ	Ə
Ɛ	Ƒ	ƒ	Ɠ	Ɣ	ƕ	Ɩ	Ɨ
Ƙ	ƙ	ƚ	ƛ	Ɯ	Ɲ	ƞ	Ɵ
Ơ	ơ	Ƣ	ƣ	Ƥ	ƥ	Ʀ	Ƨ
ƨ	Ʃ	ƪ	ƫ	Ƭ	ƭ	Ʈ	Ư
ư	Ʊ	Ʋ	Ƴ	ƴ	Ƶ	ƶ	Ʒ
Ƹ	ƹ	ƺ	ƻ	Ƽ	ƽ	ƾ	ƿ
ǀ	ǁ	ǂ	ǃ	Ǆ	ǅ	ǆ	Ǉ
ǈ	ǉ	Ǌ	ǋ	ǌ	Ǎ	ǎ	Ǐ
ǐ	Ǒ	ǒ	Ǔ	ǔ	Ǖ	ǖ	Ǘ
ǘ	Ǚ	ǚ	Ǜ	ǜ	ǝ	Ǟ	ǟ
Ǡ	ǡ	Ǣ	ǣ	Ǥ	ǥ	Ǧ	ǧ
Ǩ	ǩ	Ǫ	ǫ	Ǭ	ǭ	Ǯ	ǯ
ǰ	Ǳ	ǲ	ǳ	Ǵ	ǵ	Ƕ	Ƿ
Ǹ	ǹ	Ǻ	ǻ	Ǽ	ǽ	Ǿ	ǿ
Ȁ	ȁ	Ȃ	ȃ	Ȅ	ȅ	Ȇ	ȇ

English 01B3 Tahoma LATIN CAPITAL LETTER Y WITH HOOK

and below, the official name of the character selected. Selecting a character puts it into the clipboard ready to paste.

See also: Copying a character into Concord 420

5.1.2 wildcards

Many WordSmith functions allow you a choice of wildcards:

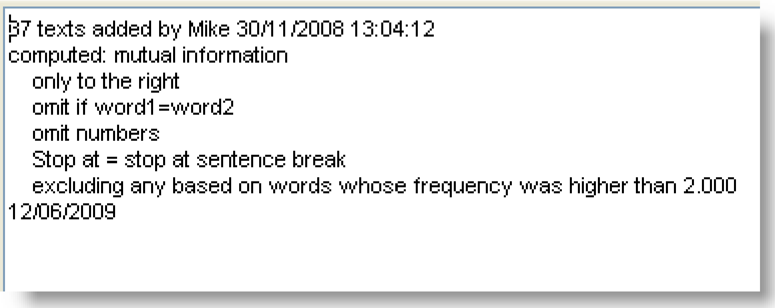
symbol	meaning	examples
*	disregard the end of the word,	tele*
	disregard a whole word	*ness
		happi

		book * hotel
?	any single character (including punctuation) will match here	**Engl???** **?50.00**
^	a single letter	**Engl^^^**
#	any sequence of numbers, 0 to 9	**$#** **£#.00**

(To represent a genuine **#,^,?** or *****, put each one in double quotes, eg. **"?" "#" "^" "*"**.)

5.2 add notes

As WordSmith generates data, it will state the current relevant settings in the Notes tab and these are saved 211 with your data. In this sample case the original work was done in 2008. In 2009, mutual information was computed on that data, with certain specific settings.



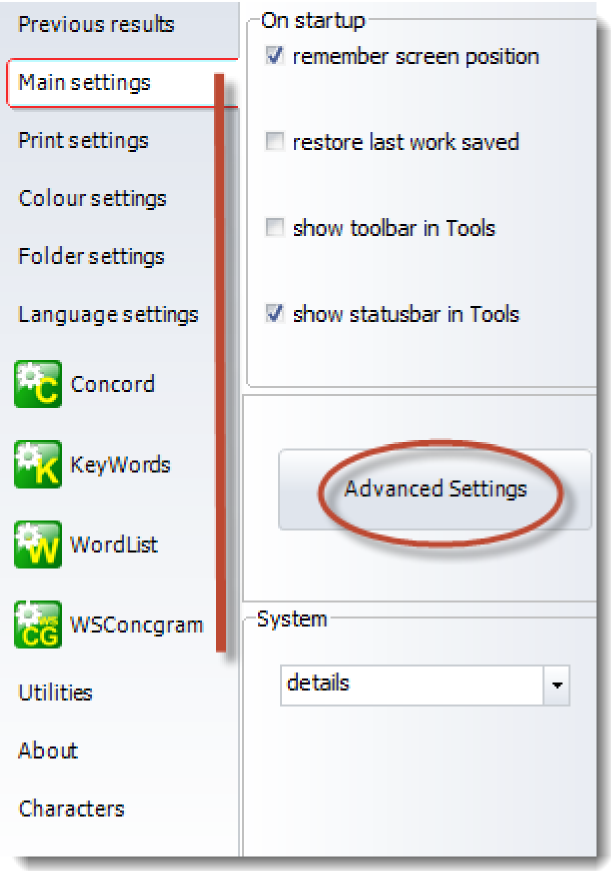
```
37 texts added by Mike 30/11/2008 13:04:12
computed: mutual information
   only to the right
   omit if word1=word2
   omit numbers
   Stop at = stop at sentence break
   excluding any based on words whose frequency was higher than 2.000
12/06/2009
```

You may add to these notes, of course. For example, if you have done a concordance and sorted it carefully using your own user-defined categories 168, you will probably want to list these and save the information for later use.

If you need access to these notes outside WordSmith Tools, select the text using Shift and the cursor arrows or the mouse, then copy it to the clipboard 422 using Ctrl+C and paste into a word processor such as notepad.

5.3 adjust settings

There are a number of Settings windows in the Controller 4. You will see tabs accessing them at the left in he main Controller window.



On startup

☑ remember screen position

☐ restore last work saved

☐ show toolbar in Tools

☑ show statusbar in Tools

System

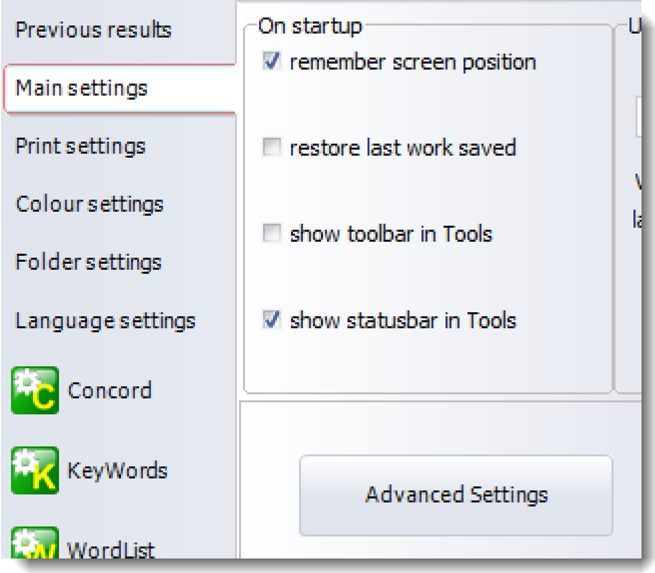
details

Choose and save 113 settings concerning:

- font 78
- colours 60
- folders 431
- tags 131
- general settings 80
- match-lists 92
- stop lists 120
- lemma lists 270
- text and language settings 124
- Concord Settings 222
- KeyWords settings 254
- WordList settings 311
- advanced user specific settings 31
- index file settings 276

5.4 advanced settings

These are reached by clicking the *Advanced Settings* button visible in the Main settings page:



Previous results

Main settings

Print settings

Colour settings

Folder settings

Language settings

Concord

KeyWords

WordList

On startup

☑ remember screen position

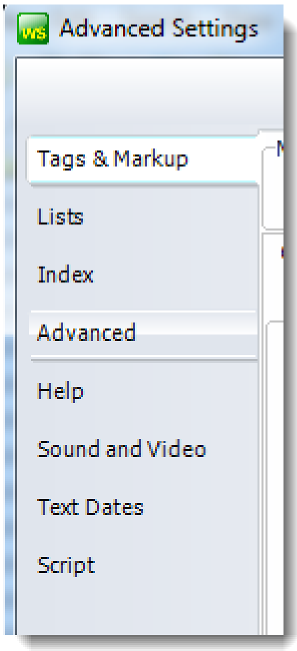
□ restore last work saved

□ show toolbar in Tools

☑ show statusbar in Tools

Advanced Settings

and open up a whole new set of options



Advanced Settings

Tags & Markup

Lists

Index

Advanced

Help

Sound and Video

Text Dates

Script

Tags & Markup 131

Lists 306

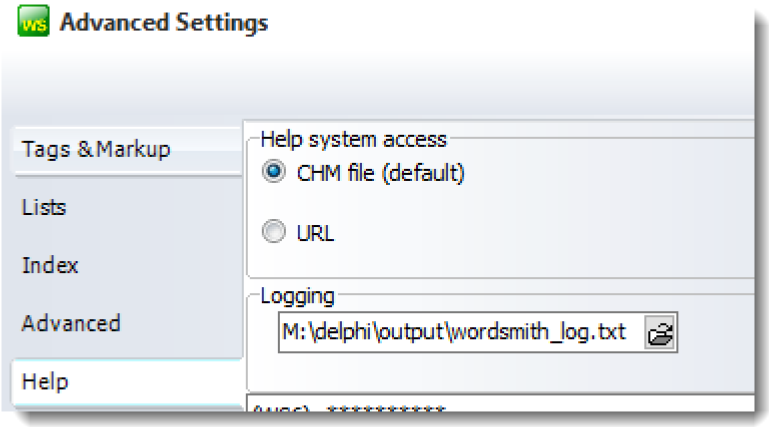
Index 310

Scripts 106

Help, logging

Help system access

On a network, it is commonly the case that Microsoft protects users to such an extent that the usual **`.CHM`** help files show only their table of contents but no details. Here you can set the WordSmith help to access the local CHM file or the online Help at the WordSmith URL.



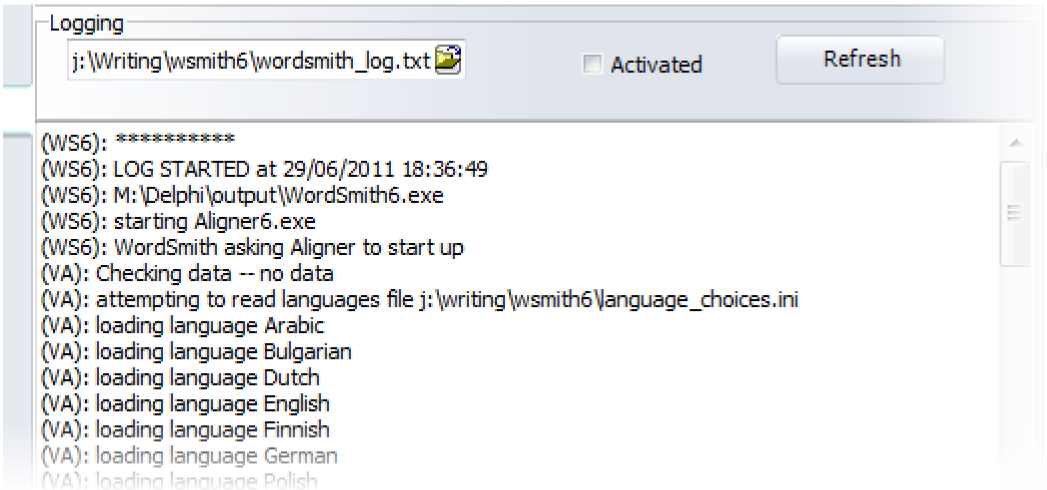
● CHM file (default)
○ URL

Logging: M:\delphi\output\wordsmith_log.txt

Logging

Logging is useful if you are getting strange results and wish to see details of how they were obtained. If this is enabled, WordSmith will save some idea of how your results are progressing in the log-file, which you see in the *Advanced Settings* | *Help* | *Logging* section in the Controller.

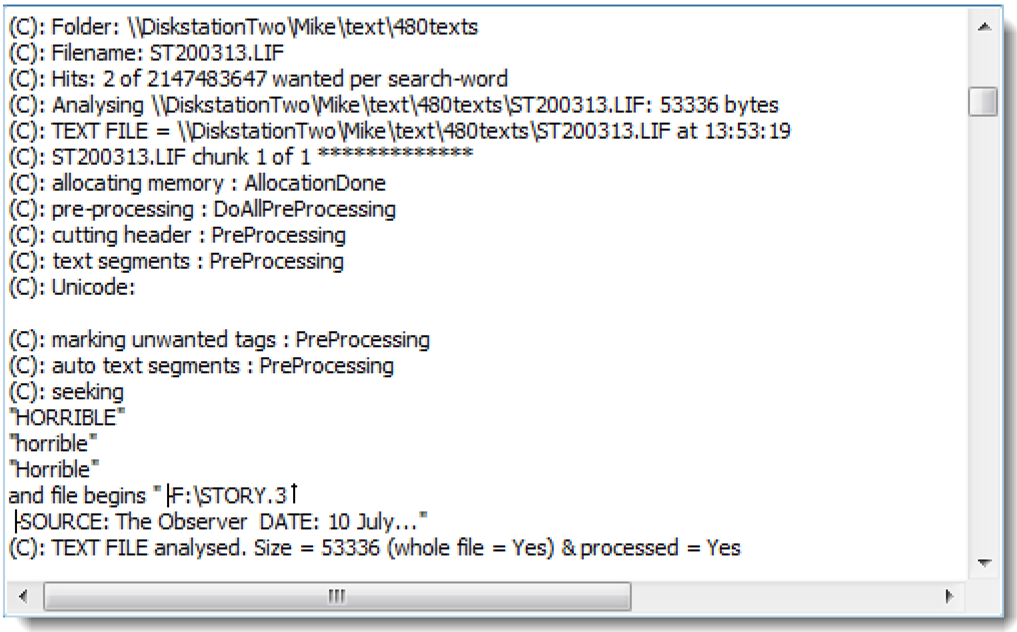
Here you can optionally switch on or off logging and choose an appropriate file-name. If you switch it on at any time you will get a chance to clear the previous log-file. This log shows WordSmith working with the Aligner, at the stage where various languages are being loaded up.



Logging: j:\Writing\wsmith6\wordsmith_log.txt □ Activated

```
(WS6): ***********
(WS6): LOG STARTED at 29/06/2011 18:36:49
(WS6): M:\Delphi\output\WordSmith6.exe
(WS6): starting Aligner6.exe
(WS6): WordSmith asking Aligner to start up
(VA): Checking data -- no data
(VA): attempting to read languages file j:\writing\wsmith6\language_choices.ini
(VA): loading language Arabic
(VA): loading language Bulgarian
(VA): loading language Dutch
(VA): loading language English
(VA): loading language Finnish
(VA): loading language German
(VA): loading language Polish
```

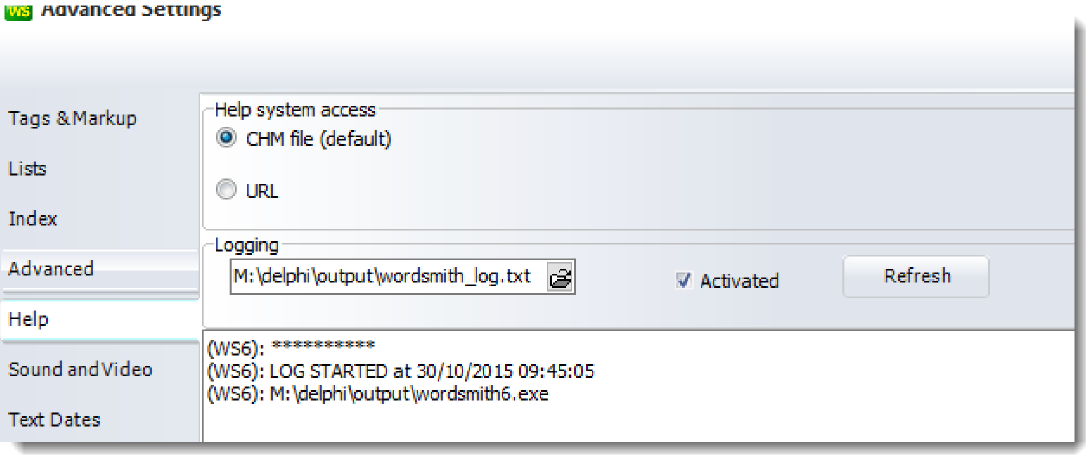
And here in a Concord process we see some details of the text files being read and processed, seeking the search-word **`horrible`**:



```
(C): Folder: \\DiskstationTwo\Mike\text\480texts
(C): Filename: ST200313.LIF
(C): Hits: 2 of 2147483647 wanted per search-word
(C): Analysing \\DiskstationTwo\Mike\text\480texts\ST200313.LIF: 53336 bytes
(C): TEXT FILE = \\DiskstationTwo\Mike\text\480texts\ST200313.LIF at 13:53:19
(C): ST200313.LIF chunk 1 of 1 **************
(C): allocating memory : AllocationDone
(C): pre-processing : DoAllPreProcessing
(C): cutting header : PreProcessing
(C): text segments : PreProcessing
(C): Unicode:

(C): marking unwanted tags : PreProcessing
(C): auto text segments : PreProcessing
(C): seeking
"HORRIBLE"
"horrible"
"Horrible"
and file begins " |F:\STORY.3↑
 |SOURCE: The Observer  DATE: 10 July..."
(C): TEXT FILE analysed. Size = 53336 (whole file = Yes) & processed = Yes
```

The most straightforward way to use logging is

1. Find logging in the Help tab of Advanced settings.



```
(WS6): ***********
(WS6): LOG STARTED at 30/10/2015 09:45:05
(WS6): M:\delphi\output\wordsmith6.exe
```

2. Click the Activated box. You'll be asked whether you want any previous log cleared.

3. Carry on using WordSmith as desired, changing settings or using Concord or any other tool. From time to time or after WordSmith finishes, press the Refresh button visible above and read the output. It is a text file so it can be opened using any word processing software. If you have had trouble, looking at the last few lines may help by showing where processing stopped.

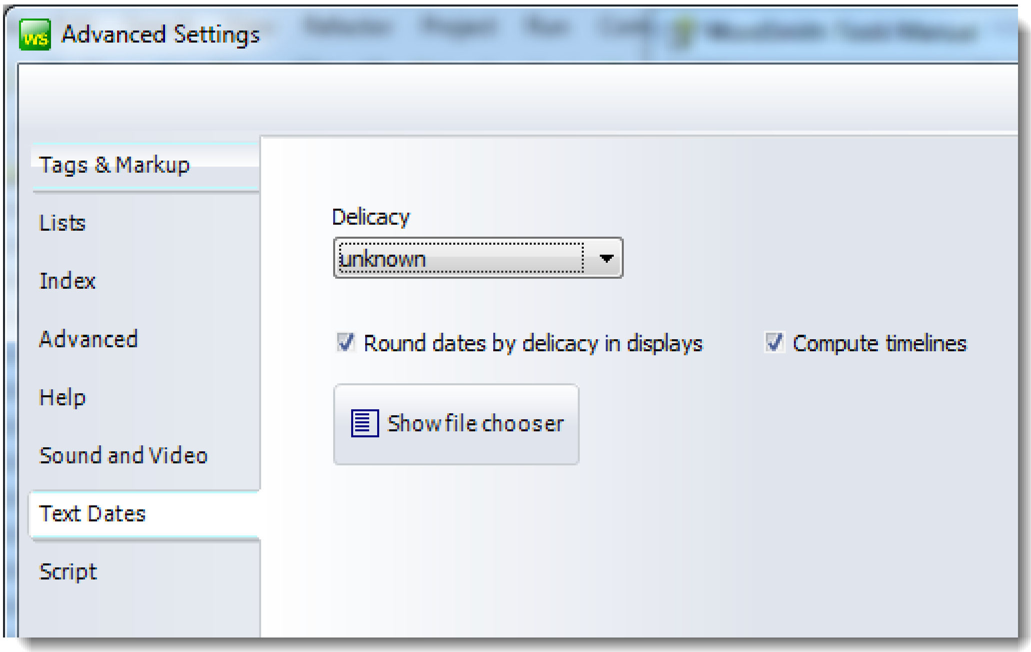
If you want to log as WordSmith starts up, start in from the command line with the parameter **`/log`**:

`Start | Run | Cmd` <Enter> **`| cd\wsmith6`** <Enter> **`| wordsmith6 /log`** <Enter>
(or **`wordsmith6 /log C:\temp\WSLog.txt`** to force use of **`C:\temp\WSLog.txt`**. If you do that, make sure the folder exists first.)

See also: emailed error reports 417.

Text Dates

Text dates can be set to varying levels of delicacy, depending on the range of text file dates chosen.



Delicacy: unknown

☑ Round dates by delicacy in displays

☑ Compute timelines

See also: using text dates 126

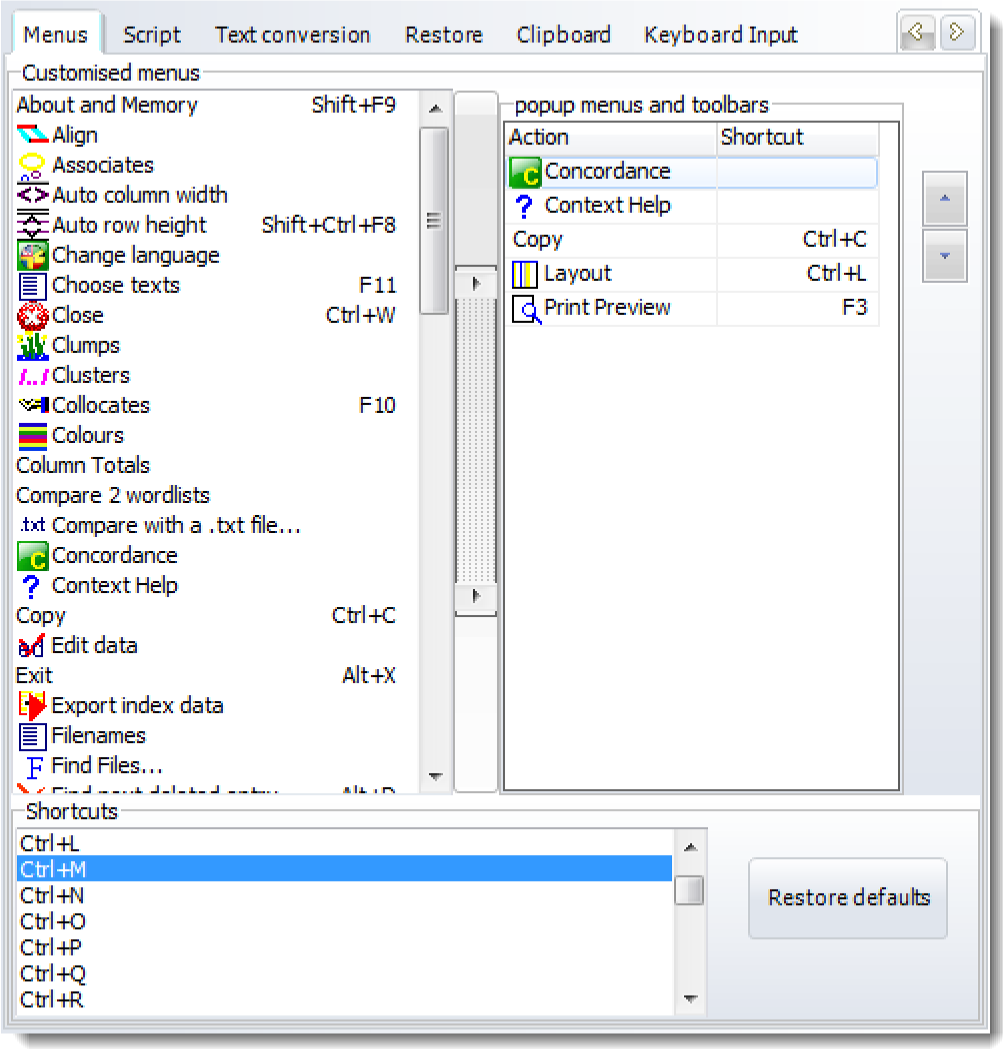
Advanced section (menus, clipboard, deadkeys etc.)

Customising menus

You can re-assign new shortcuts (such as Alt+F3, Ctrl+O) to the menu items 441 which are used in the various Tools.

And all grids of data have a "popup menu" which appears when you click the right button of your mouse.

To customise this, in the main WordSmith Controller program, choose *Main Settings* | *Advanced* | *Menus*.



Customised menus	
About and Memory	Shift+F9
Align	
Associates	
Auto column width	
Auto row height	Shift+Ctrl+F8
Change language	
Choose texts	F11
Close	Ctrl+W
Clumps	
Clusters	
Collocates	F10
Colours	
Column Totals	
Compare 2 wordlists	
Compare with a .txt file...	
Concordance	
Context Help	
Copy	Ctrl+C
Edit data	
Exit	Alt+X
Export index data	
Filenames	
Find Files...	

Action	Shortcut
Concordance	
Context Help	
Copy	Ctrl+C
Layout	Ctrl+L
Print Preview	F3

Shortcuts:
Ctrl+L
Ctrl+M
Ctrl+N
Ctrl+O
Ctrl+P
Ctrl+Q
Ctrl+R

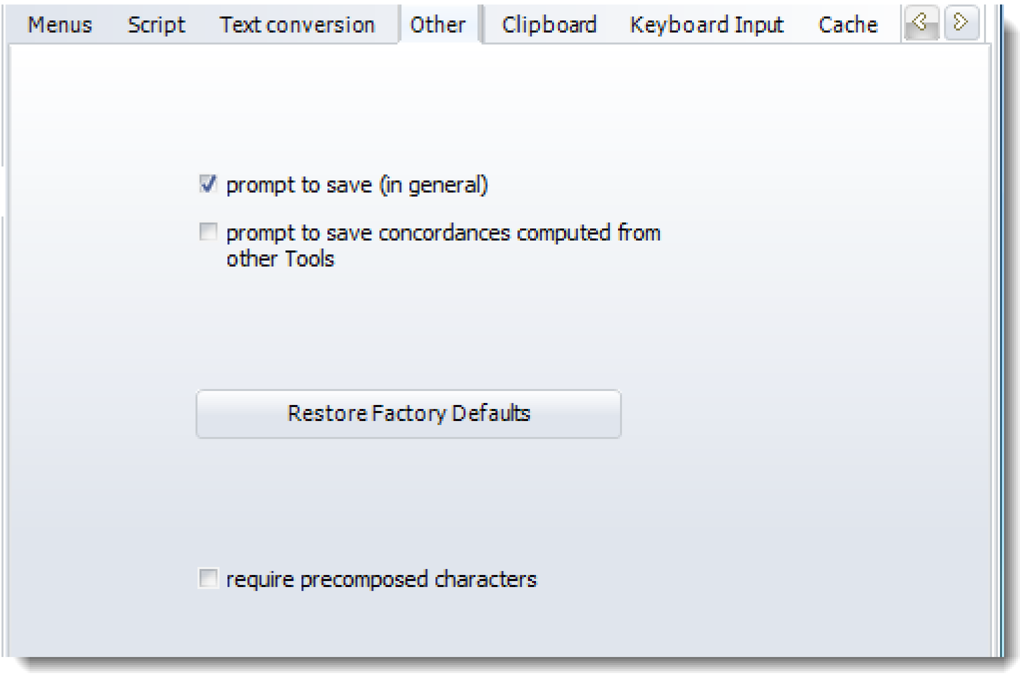
You will see a list of menu options at the left, and can add to (or remove from) the list on the right by selecting one on the left and pressing the buttons in the middle, or by dragging it to the right. To re-order the choices, press the up or down arrow. In the screenshot I've added "Concordance" as I usually want to generate concordances from word-lists and key word lists.

Whatever is in your popup menu will also appear in the Toolbar 80.

Below, you see a list of Shortcuts, with Ctrl+M selected. To change a shortcut, drag it up to the Customised menu list or the popup menus and toolbars list.

The Restore defaults button puts all shortkeys back to factory settings. To save the choices permanently, see Saving Defaults 113.

Other



☑ prompt to save (in general)

□ prompt to save concordances computed from other Tools

□ require precomposed characters

Here you may press a button to restore all factory defaults, useful if your settings are giving trouble.

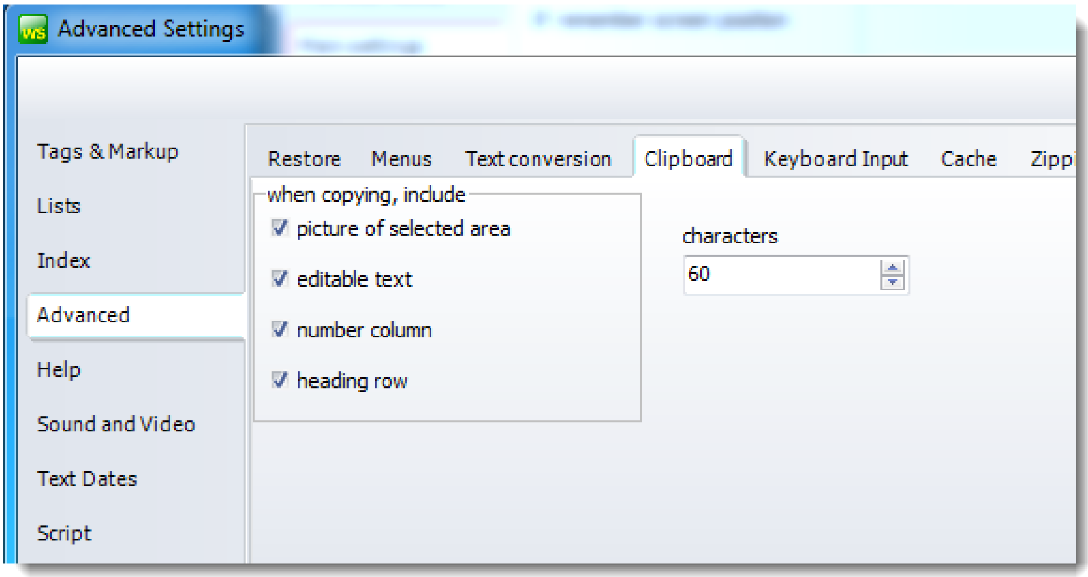
prompt to save (in general): reminds you to save every time new data results are computed or re-organised.

prompt to save concordances computed from other Tools: (default=false) prompt after WordList or KeyWords or WSConcGram gets a concordance computed.

require precomposed characters: some languages have a lot of cases where two characters get merged in the display into one, e.g. **e** with ` appearing as **è**. WordSmith will automatically check for such pairs when processing languages such as Amharic, Arabic, Bengali, Farsi, Gujarati, Hindi, Kannada, Khmer, Lao, Malayalam, Nepali, Oriya, Thai, Tibetan, Telegu, Tamil, Yoruba. If you want to force WordSmith to carry out a test for such pairs when processing all languages, however, check this box.

Clipboard

Here you may choose defaults for copying.



Advanced Settings

- Tags & Markup
- Lists
- Index
- Advanced
- Help
- Sound and Video
- Text Dates
- Script

Restore | Menus | Text conversion | Clipboard | Keyboard Input | Cache | Zippi

when copying, include

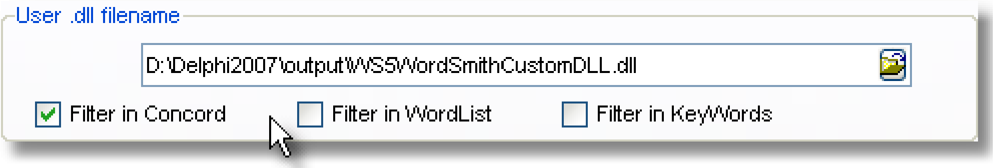
- ☑ picture of selected area
- ☑ editable text
- ☑ number column
- ☑ heading row

characters: 60

The number of characters only applies when copying as editable text. See also: clipboard 422

User .dll

If you have a DLL which you want to use to intercept WordSmith's results, you can choose it here. The one this user is choosing, **`WordSmithCustomDLL.dll`**, is supplied with your installation and can be used when you wish. If "Filter in Concord" is checked, this .dll will append all concordance lines found in plain text to a file called **`Concord_user_dll_concordance_lines.txt`** in your **`\wsmith6`** folder, if there is space on the hard disk.



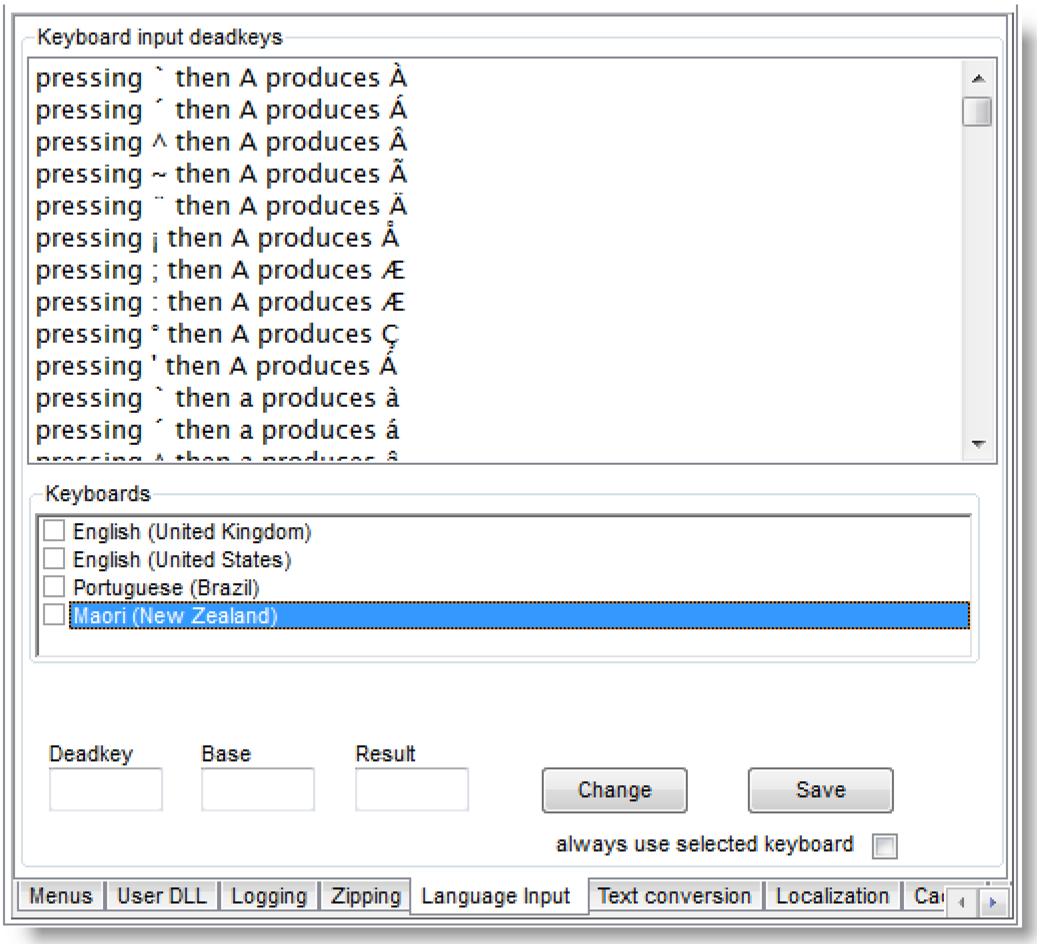
User .dll filename

D:\Delphi2007\output\WS5\WordSmithCustomDLL.dll

☑ Filter in Concord　☐ Filter in WordList　☐ Filter in KeyWords

Language Input

Deadkeys are used to help type accented characters with some keyboards. The language input tab lets you alter the deadkeys to suit your keyboard and if necessary force WordSmith to use the keyboard layout of your choice whenever WordSmith starts up.



Keyboard input deadkeys

pressing ` then A produces À
pressing ´ then A produces Á
pressing ^ then A produces Â
pressing ~ then A produces Ã
pressing ¨ then A produces Ä
pressing ¡ then A produces Å
pressing ; then A produces Æ
pressing : then A produces Æ
pressing ¸ then A produces Ç
pressing ' then A produces Á
pressing ` then a produces à
pressing ´ then a produces á

Keyboards

□ English (United Kingdom)
□ English (United States)
□ Portuguese (Brazil)
□ Maori (New Zealand)

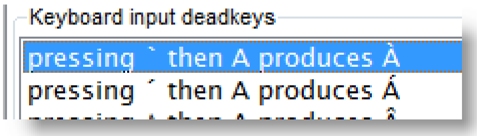
Deadkey | Base | Result

always use selected keyboard □

Here the user's Windows has four keyboard layouts installed. To type in Maori, you might choose to select Maori, and change a couple of deadkeys. At present, as the list shows, pressing ` then **A** gives **À**, but users of Maori usually prefer that combination to give **Ā**.

To change these settings,

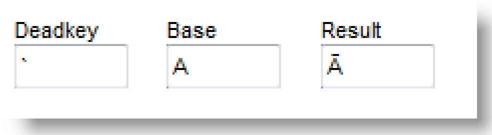
1. select the line



Keyboard input deadkeys

pressing ` then A produces À
pressing ´ then A produces Á

2. edit the box below:



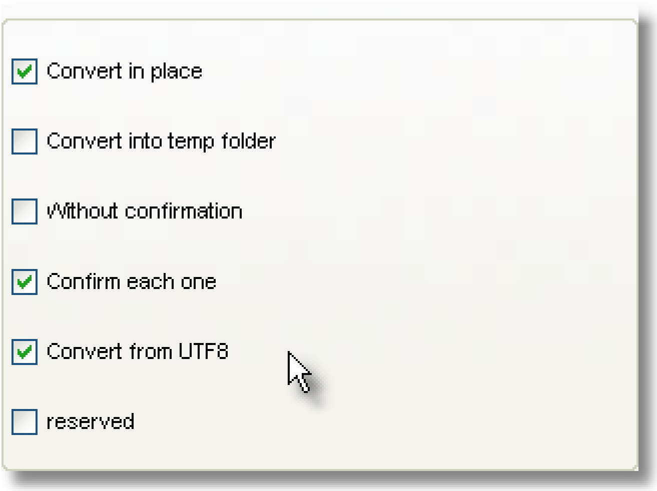
Deadkey: `
Base: A
Result: Ā

(you can drag the character you need from the [character window](28))

then press Change. When you've changed all the characters you want, press Save. If you want WordSmith to force the keyboard to Maori too every time it starts (this will probably be necessary if it is not a New Zealand computer) then check the *always use selected keyboard* box.

Text Conversion

If your text files happen to contain UTF-8 text files, WordSmith will notice and may offer to convert them on the spot using the options below.



☑ Convert in place

□ Convert into temp folder

□ Without confirmation

☑ Confirm each one

☑ Convert from UTF8

□ reserved

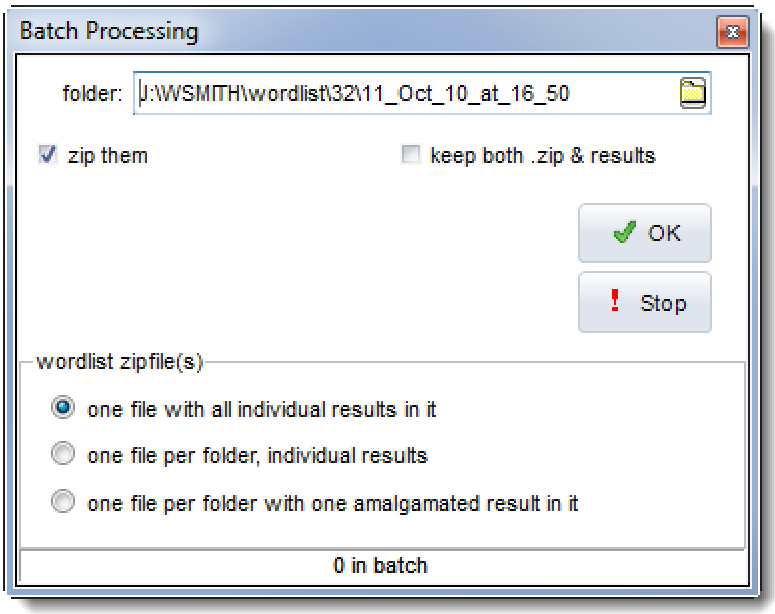
See also : menu and button options 441.

5.5 batch processing

The point of it...

Batch processing is used when you want to make separate lists, but you don't want the trouble of doing it one by one, manually selecting each text file, making the word list or concordance, saving it, and so on.

If you have selected more than one text file you can ask WordList, Concord and KeyWords to process as a batch.



Batch Processing

folder: J:\WSMITH\wordlist\32\11_Oct_10_at_16_50

☑ zip them

□ keep both .zip & results

wordlist zipfile(s)

● one file with all individual results in it

○ one file per folder, individual results

○ one file per folder with one amalgamated result in it

0 in batch

Folder where they end up

The name suggested is today's date[425]. Edit it if you like. Whatever you choose will get created when the batch process starts.

The results will be stored in folders stemming from the folder name. That is, if you start making word lists in

`c:\wsmith\wordlist\05_07_19_12_00,` they will end up like this:
`c:\wsmith\wordlist\05_07_19_12_00\0\fred1.lst`
`c:\wsmith\wordlist\05_07_19_12_00\0\jim2.lst`
`..`
`c:\wsmith\wordlist\05_07_19_12_00\0\mary512.lst`
then
`c:\wsmith\wordlist\05_07_19_12_00\1\joanna513.lst`
etc.

Filenames will be the source text filename with the standard extension **`(.lst, .cnc, .kws)`**.

Zip them

If checked, the results are physically stored in a standard **`.zip`** file. You can extract them using your standard zipping tool such as Winzip, or you can let WordSmith do it for you. The files within are exactly the same as the uncompressed versions but save disk space -- and the disk system will also be less unhappy than if there are many hundreds of files in the same folder.

If you zip them, you will get

`c:\wsmith\wordlist\05_07_19_12_00\batch.zip`

and all the sub-files will get deleted unless you check "keep both .zip and results".

One file / One file per folder?

The first alternative (default) makes one .zip file with all your individual word-lists in it. Each word-list or concordance or keywords list is for one source text.

But what if your text files are structured like this:

`\..\BNC`
`\..\BNC\written`

```
\..\BNC\written\humanities
\..\BNC\written\medicine
\..\BNC\written\science
\..\BNC\spoken
```

etc.

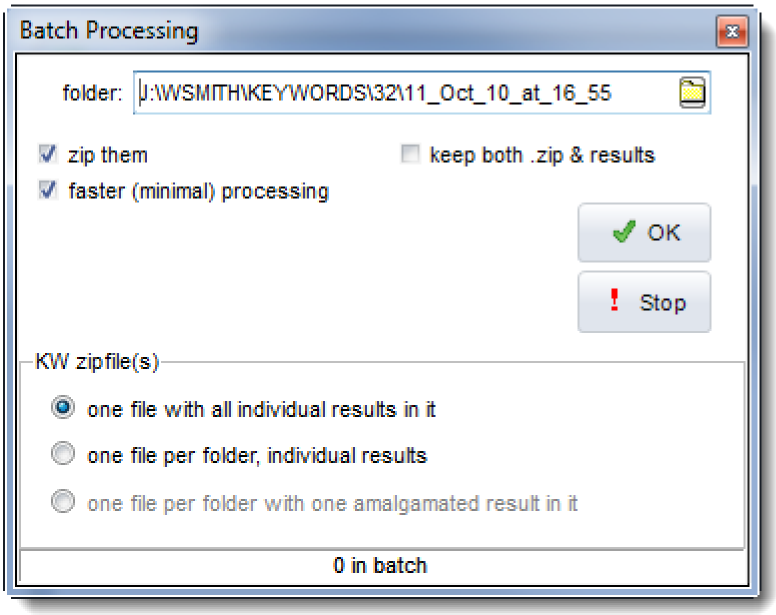
The *One file per folder, individual zipfiles* makes a separate .zip of each separate folderful of textfiles (eg. one for humanities, another for medicine, etc.), with one list for each source text.

The *One file per folder, amalgamated zipfiles* makes a separate .zip of each folderful, but makes one word-list or concordance from that whole folderful of texts.

Batch Processing and Excel

These options may also offer a chance for data to be copied automatically to an Excel file.

Faster (Minimal) Processing



Batch Processing

folder: J:\WSMITH\KEYWORDS\32\11_Oct_10_at_16_55

☑ zip them

☐ keep both .zip & results

☑ faster (minimal) processing

KW zipfile(s)

● one file with all individual results in it

○ one file per folder, individual results

○ one file per folder with one amalgamated result in it

0 in batch

This checkbox is only enabled if you are about to start a process where more than one kind of result can be computed simultaneously. For example, if you are computing a concordance, by default collocates 179, patterns 207 and dispersion plots 191 will be computed when each concordance is done. In KeyWords, likewise, there will be dispersion plots 251, link 247 calculations etc. which will be computed as the KWs are calculated.

If checked, only the minimal computation will be done (KWs in *KeyWords* processing, concordance in *Concord*). This will be faster, and you can always get the plots computed later as long as the source texts 430 don't get moved or deleted.

Example: you're making word lists and have chosen 1,200 text files which are from a magazine called "The Elephant".

You specify

`C:\WSMITH\WORDLIST\ELEPHANT` as your folder name.

If you already have a folder called `C:\WSMITH\WORDLIST\ELEPHANT`, you will be asked for permission to erase it and all sub-folders of it!

After you press OK,

1,200 new word-lists are created, called `trunk.LST, tail.LST .. digestive system.LST`. They are all in numbered sub-folders of a folder called `C:\WSMITH\WORDLIST\ELEPHANT`.

If you did not check "zip them into 1 .zip file", you will find them under `C:\WSMITH\WORDLIST\ELEPHANT\0`.

If you did check "zip them into 1 .zip file", there is now a `C:\WSMITH\WORDLIST\ELEPHANT.ZIP` file which contains all your results. (The 1,200 `.LST` files created will have been erased but the `.ZIP` file contains all your lists.)

The advantage of a `.zip` file is that it takes up much less disk space and is easy to email to others. WordSmith can access the results from within a `.zip` file, letting you choose which word list, concordance etc. you want to see.

Getting at the results in WordSmith

Choose *File* | *Open* as usual, then change the file-type to "Batch file *.zip". When you choose a .zip file, you will see a window listing its contents. Double-click on any one to open it.

Note: of course Concord will only succeed in opening a concordance and KeyWords a key word list file. If you choose a .zip file which contains something else, it will give an error message.

See also: batch scripts 106

5.6 choosing texts

This chapter explains how to select texts, save a selection and even attach a date going back as far as 4000BC to each text file.

You need text in a suitable format 42.

5.6.1 text formats

In WordSmith you need plain text files 419, such as you get if you save a Word .doc 444 as Plain Text (**`.txt`**). The text format should be ASCII or ANSI or Unicode (UTF16).

Any Word **`.doc`** or **`.docx`** files will look crossed out and should not be used: convert them to .txt first 444.

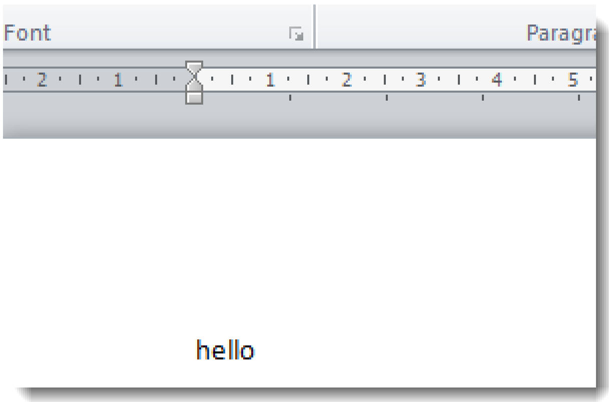
Text files within .zip files 50 can be used; they will be coloured red in the *Files available* display but WordSmith can read them and get the texts you select within them.

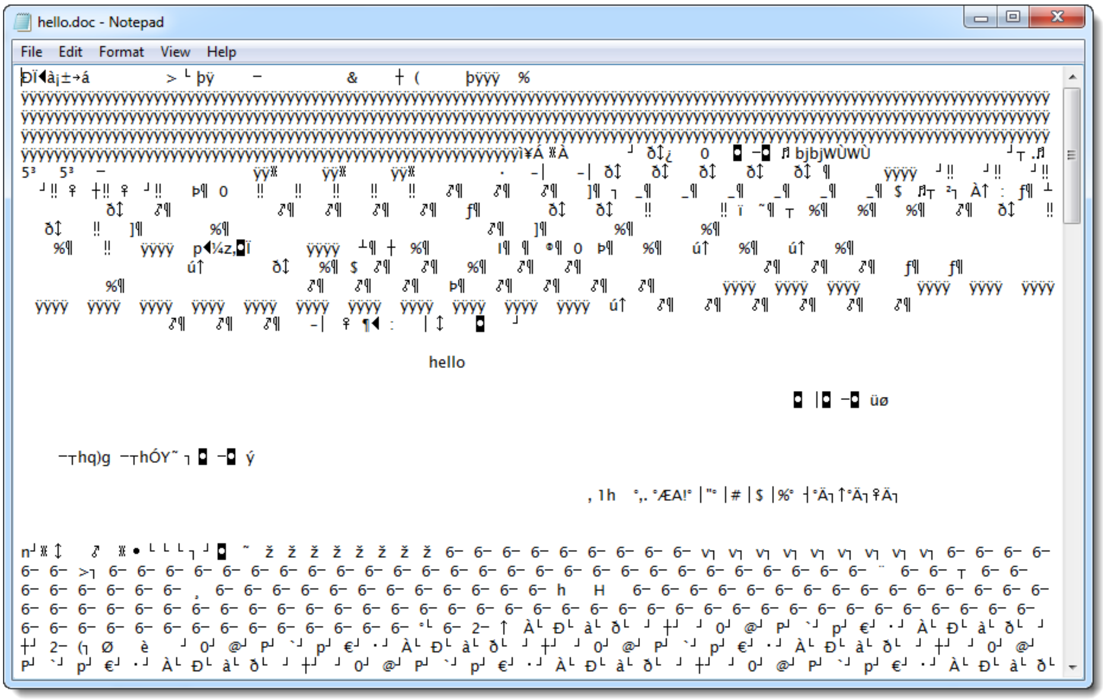
Why not .PDF files?

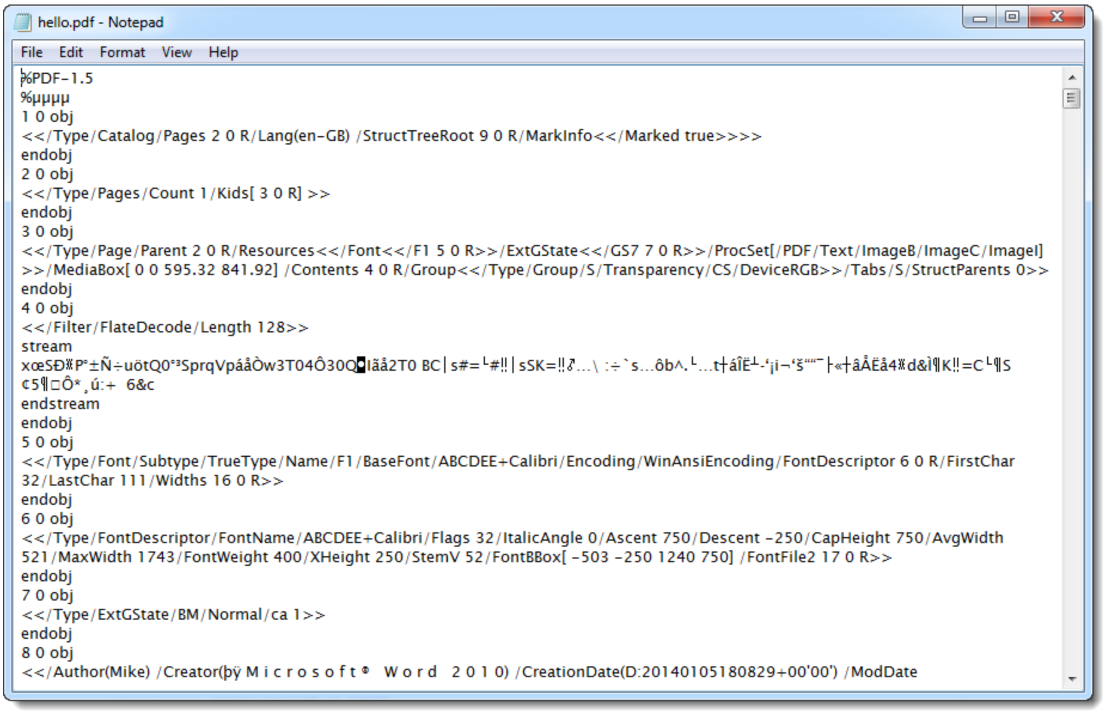
Don't choose **`.pdfs`** either, they have a very special format. Essentially a PDF is a set of instructions telling a printer or browser where to place coloured dots. The plain text is usually hard to extract even if you use Adobe Acrobat (and Adobe invented the format).

Why not .DOC files?

A **`.DOC`** is rather unsuitable even if it does contain the text: this is what a **`.DOC`** containing only the word **`hello`** looks like in Word, then opened up in Notepad, then the **`.PDF`** of the same.







```
%PDF-1.5
%µµµµ
1 0 obj
<</Type/Catalog/Pages 2 0 R/Lang(en-GB) /StructTreeRoot 9 0 R/MarkInfo<</Marked true>>>>
endobj
2 0 obj
<</Type/Pages/Count 1/Kids[ 3 0 R] >>
endobj
3 0 obj
<</Type/Page/Parent 2 0 R/Resources<</Font<</F1 5 0 R>>/ExtGState<</GS7 7 0 R>>/ProcSet[/PDF/Text/ImageB/ImageC/ImageI]
>>/MediaBox[ 0 0 595.32 841.92] /Contents 4 0 R/Group<</Type/Group/S/Transparency/CS/DeviceRGB>>/Tabs/S/StructParents 0>>
endobj
4 0 obj
<</Filter/FlateDecode/Length 128>>
stream
xœSÐ‼P°±Ñ÷uötQ0°³SprqVpáåÒw3T04Ô30Q█Iãå2T0 BC|s#=└#‼|sSK=‼♂...\ :÷`s...ôb^.└...t┼áÎË┴-'¡i¬'š""¯├«┼âÅËå4‼d&Ì¶K‼=C└¶S
¢5¶□Ô*¸ú:+  6&c
endstream
endobj
5 0 obj
<</Type/Font/Subtype/TrueType/Name/F1/BaseFont/ABCDEE+Calibri/Encoding/WinAnsiEncoding/FontDescriptor 6 0 R/FirstChar
32/LastChar 111/Widths 16 0 R>>
endobj
6 0 obj
<</Type/FontDescriptor/FontName/ABCDEE+Calibri/Flags 32/ItalicAngle 0/Ascent 750/Descent -250/CapHeight 750/AvgWidth
521/MaxWidth 1743/FontWeight 400/XHeight 250/StemV 52/FontBBox[ -503 -250 1240 750] /FontFile2 17 0 R>>
endobj
7 0 obj
<</Type/ExtGState/BM/Normal/ca 1>>
endobj
8 0 obj
<</Author(Mike) /Creator(þÿ M i c r o s o f t ®   W o r d   2 0 1 0) /CreationDate(D:20140105180829+00'00') /ModDate
```

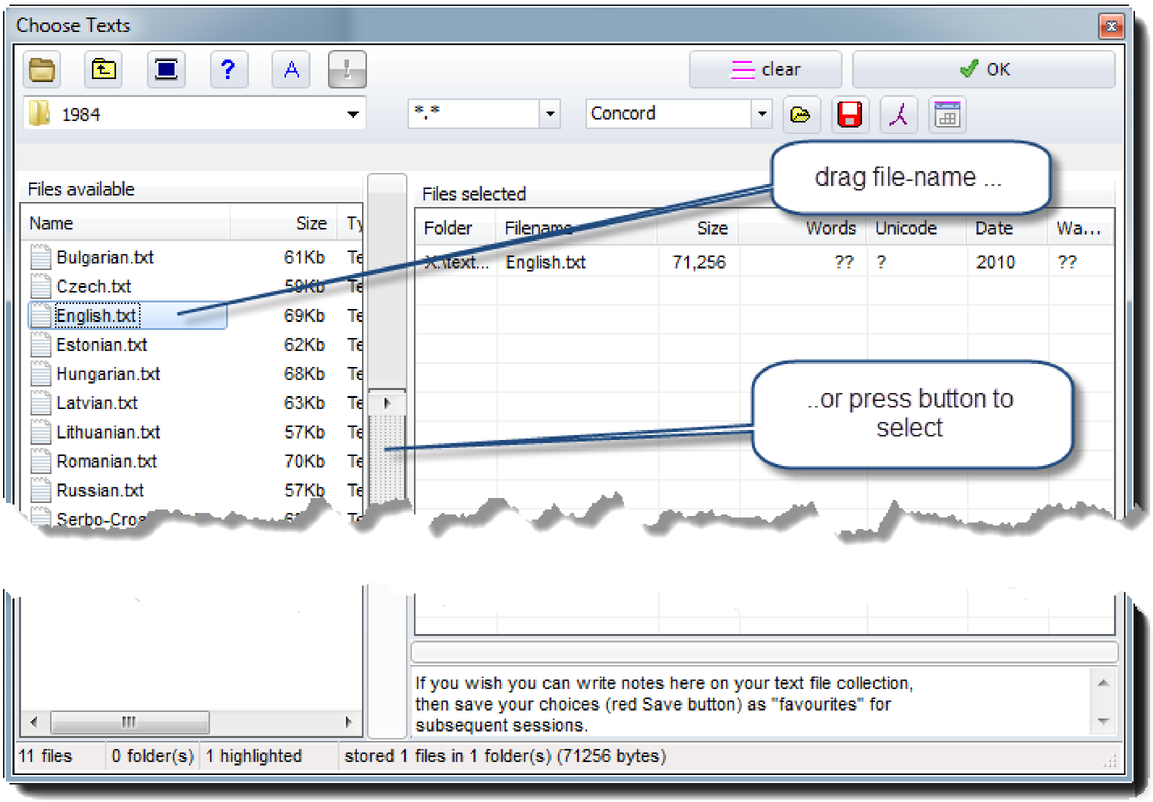
Check the format is OK

In the file-choose window you can test the format of the texts you've chosen with the *Test File Format* 46 () button.

5.6.2 the file-choose window

How to get here

This function is accessed from the File menu in the Controller 4 and the Settings menu or New menu item (●) in the various Tools.



Files available

Name	Size	Ty
Bulgarian.txt	61Kb	Te
Czech.txt	59Kb	Te
English.txt	69Kb	Te
Estonian.txt	62Kb	Te
Hungarian.txt	68Kb	Te
Latvian.txt	63Kb	Te
Lithuanian.txt	57Kb	Te
Romanian.txt	70Kb	Te
Russian.txt	57Kb	Te
Serbo-Cro	6	Te

Files selected

Folder	Filename	Size	Words	Unicode	Date	Wa...
X:\text...	English.txt	71,256	??	?	2010	??

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.

11 files | 0 folder(s) | 1 highlighted | stored 1 files in 1 folder(s) (71256 bytes)

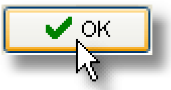
The two main areas at left and right are

- files to choose from (at left)
- files already selected (at right)

The button which the red arrow points at is what you press to move any you have selected at the left to your "files selected" at the right. Or just drag them from the left to the right.

The list on the right shows full file details (name, date, size, number of words (above shown with ?? as WordSmith doesn't yet know, though it will after you have concordanced or made a word list) and whether the text is in Unicode (? for the same reason). To the right of Unicode is a column stating whether each text file meets your requirements 137.

If you have never used WordSmith before (more precisely if you have not yet saved any concordances, word lists etc.) you will find that a chapter from Charles Dickens' Tale of 2 Cities has been selected for you. To stop this happening, make sure that you do save at least one word list or concordance! See also -- previous lists 97.



This puts the current file selection into store. All files of the type you've specified in any sub-folders will also get selected if the "Sub-folders too" checkbox is checked. You can check on which ones have been selected under All Current Settings.

Clear

As its name suggests, this allows you to change your mind and start afresh. If any selected filenames are highlighted, only these will be cleared.

More details

File Types

The default file specification is *.* (i.e. any file) but this can be altered in the box or set permanently in wordsmith6.ini 113.

Tool

In the screenshot above you can see Concord -- we are choosing texts for Concord. There are alternatives available (WordList, KeyWords etc.).

Select All

Selects all the files in the current folder.

Drives and Folders

Double-click on a folder to enter it. You can re-visit a folder if its name is in the folder window history list, and easily go back with the standard Windows "back" button. Or click on the button to choose a new drive or folder.

Sub-Folders

If checked, when you select a whole driveful or a whole folderful of texts at the left, you will select it plus any files in any sub-folders of that drive or folder.

Sorting

By clicking on the column headers (*Folder, Filename, Size, Type, Words, Unicode, Date* etc.) you can re-sort the listing.

Test text format

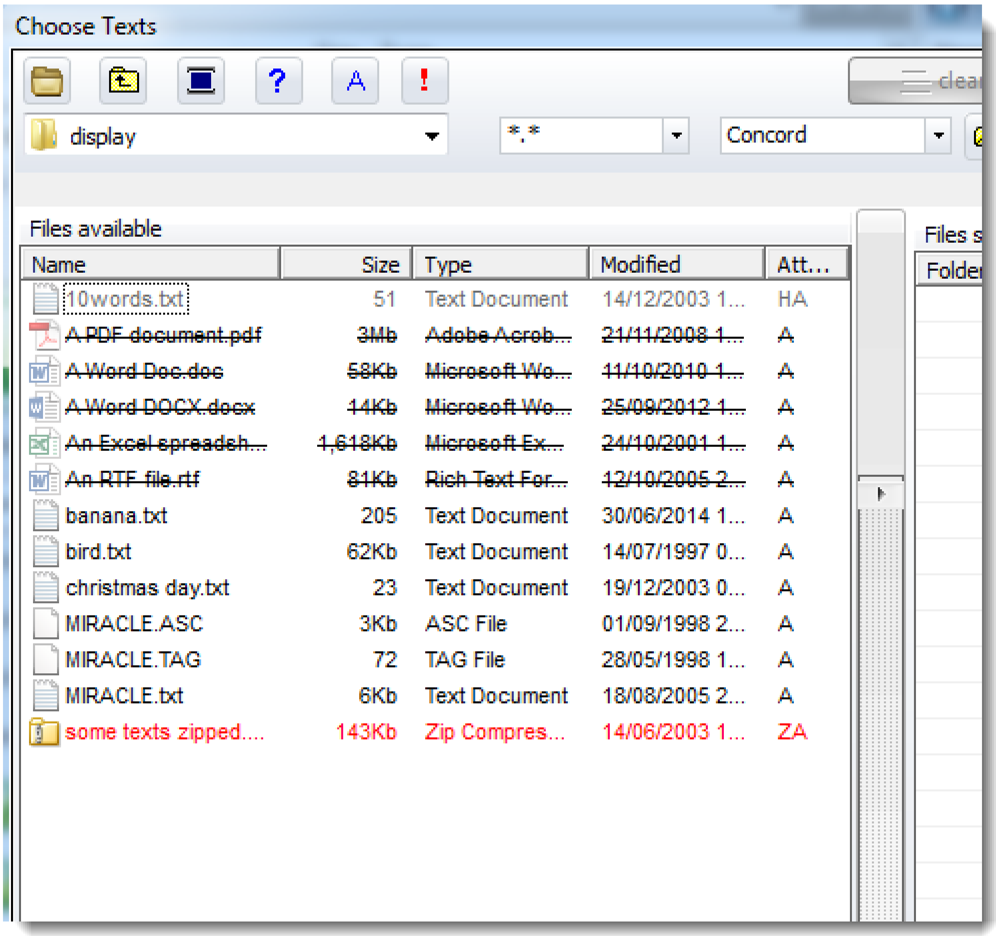
This button checks the format of any files selected. In the screenshot above, no tests have been done so the display shows ? for each file. If the text file is in Unicode, the display shows *U*, if Unicode big-endian it'll show *UB*, if plain ASCII or Ansi text it will show *A*, if it's a Word .doc file, *D*. If it is in UTF-8, *8*. If you get inconsistency you'll be invited to convert them all to Unicode.

Favourites 50

Two buttons on the right (and) allow you to save or get a previous file selection 50, saving you the trouble of making and remembering a complex set of choices.

Type of text files

In WordSmith you need plain text files 419, such as you get if you save a **`Word .doc`** 444 as Plain Text (**`.txt`**). Any Word **`.doc`** files or **`.pdf`**s will look crossed out and should not be used: convert them to .txt first 444.



Name	Size	Type	Modified	Att...
10words.txt	51	Text Document	14/12/2003 1...	HA
~~A PDF document.pdf~~	~~3Mb~~	~~Adobe Acrob...~~	~~21/11/2008 1...~~	~~A~~
~~A Word Doc.doc~~	~~58Kb~~	~~Microsoft Wo...~~	~~11/10/2010 1...~~	~~A~~
~~A Word DOCX.docx~~	~~14Kb~~	~~Microsoft Wo...~~	~~25/09/2012 1...~~	~~A~~
~~An Excel spreadsh...~~	~~1,618Kb~~	~~Microsoft Ex...~~	~~24/10/2001 1...~~	~~A~~
~~An RTF file.rtf~~	~~81Kb~~	~~Rich Text For...~~	~~12/10/2005 2...~~	~~A~~
banana.txt	205	Text Document	30/06/2014 1...	A
bird.txt	62Kb	Text Document	14/07/1997 0...	A
christmas day.txt	23	Text Document	19/12/2003 0...	A
MIRACLE.ASC	3Kb	ASC File	01/09/1998 2...	A
MIRACLE.TAG	72	TAG File	28/05/1998 1...	A
MIRACLE.txt	6Kb	Text Document	18/08/2005 2...	A
some texts zipped....	143Kb	Zip Compres...	14/06/2003 1...	ZA

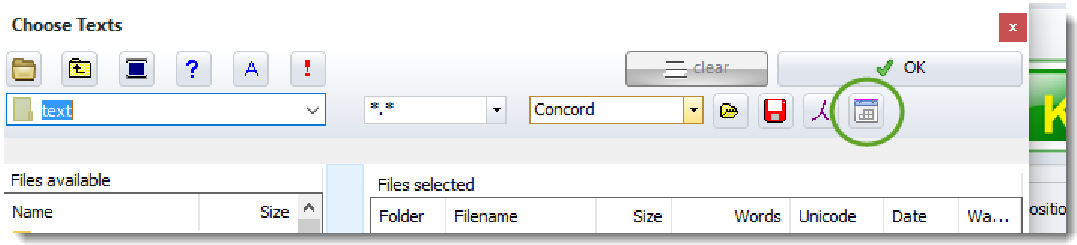
(*10words.TXT* is greyed out because it has the hidden attribute)

Setting text file dates

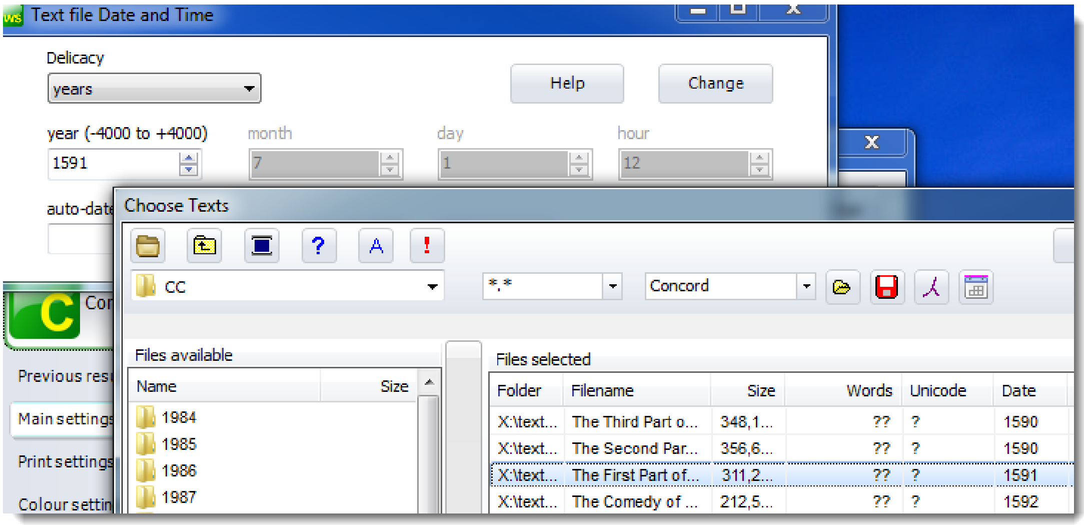
You can edit the textual date to be attached to any text file within any date range from 4000BC upwards. (On first reading from disk the date will be set to the date that text file was last edited.)

How to do it

Press the button circled in this screen-shot:



A window opens up letting you set text file dates and times. Here below you will see Shakespeare plays with their dates being edited.



Delicacy: years

year (-4000 to +4000): 1591

month: 7

day: 1

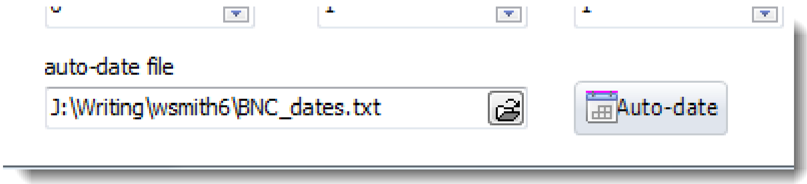
hour: 12

Name
1984
1985
1986
1987

Folder	Filename	Size	Words	Unicode	Date
X:\text...	The Third Part o...	348,1...	??	?	1590
X:\text...	The Second Par...	356,6...	??	?	1590
X:\text...	The First Part of...	311,2...	??	?	1591
X:\text...	The Comedy of ...	212,5...	??	?	1592

Delicacy offers a choice of various time ranges (centuries, years, etc.) which will help ignore excessive detail. If years are chosen as above, month, day and hour of editing are no longer relevant and default to 1st July at 12:00.

If you choose a suitable text file and press the Auto-date button, each of your chosen text files will be updated if its file-name and a suitable date are found in the list.



auto-date file: J:\Writing\wsmith6\BNC_dates.txt

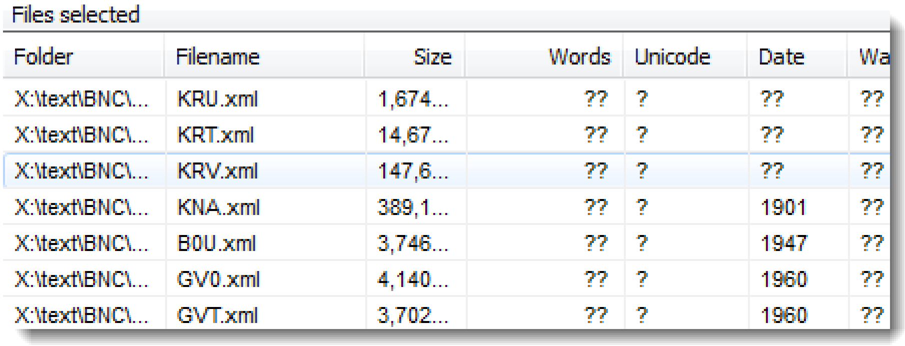
The format of the list is

`filename<tab>date` (formatted YYYY or YYYY/MM/DD for year, month and day)

Examples:

```
A0X   1991
B03   1992/04/17
```

Here we see BNC text files sorted by date. The ones at the top had no date, then the first dated was **`KNA.XML`** (a spoken sermon) dated as 1901, which is when the header says the tape-recording was made(!).



Files selected

Folder	Filename	Size	Words	Unicode	Date	Wa
X:\text\BNC\...	KRU.xml	1,674...	??	?	??	??
X:\text\BNC\...	KRT.xml	14,67...	??	?	??	??
X:\text\BNC\...	KRV.xml	147,6...	??	?	??	??
X:\text\BNC\...	KNA.xml	389,1...	??	?	1901	??
X:\text\BNC\...	B0U.xml	3,746...	??	?	1947	??
X:\text\BNC\...	GV0.xml	4,140...	??	?	1960	??
X:\text\BNC\...	GVT.xml	3,702...	??	?	1960	??

Your %USER_FOLDER% folder includes an auto-date file for the BNC (**`BNC dates`**) and another for the Shakespeare corpus (**`Shakespeare plays dated plain`**).

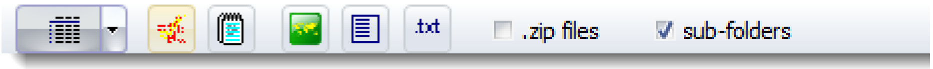
There is also a utility in the File Utilities 352 which can parse text files to generate dates using your own syntax, preparing a text file like this to read in.

You can save the dates and files as favourites 50 so as to re-use this information as often as you like.

See also: using text dates 126

Advanced

Opens a toolbar showing some further buttons:



The buttons at the top left let you see the files available as icons, as a list, or with full details (the default) instead.

Random

This re-orders the files (on both sides) in random order.

View in Notepad

Lets you see the text contents in the standard Windows simple word-processor for text files, Notepad.

Get from Internet

Allows you to access WebGetter 12 so as to download text from the Internet.

Check

Checks whether the files selected are available to read (e.g. after loading up Favourites).

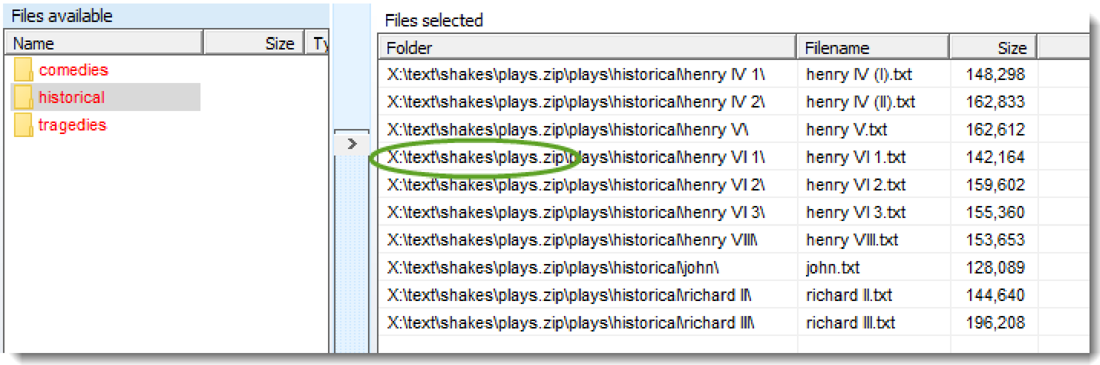
Save List

Lets you save any already stored text files as a plain text list (e.g for adding date information).

Zip files

If checked, when loading up a whole folder of text files, WordSmith will automatically include ones from .zip files.

Whether checked or not, if you double-click on a zip file 453 you can enter that as if it were a folder and see the contents. Zip files will be coloured red. In this screen, the historical plays of Shakespeare within a zip file (**`plays.zip`**) have been selected.



Files available

Name	Size	Ty
comedies		
historical		
tragedies		

Files selected

Folder	Filename	Size
X:\text\shakes\plays.zip\plays\historical\henry IV 1\	henry IV (I).txt	148,298
X:\text\shakes\plays.zip\plays\historical\henry IV 2\	henry IV (II).txt	162,833
X:\text\shakes\plays.zip\plays\historical\henry V\	henry V.txt	162,612
X:\text\shakes\plays.zip\plays\historical\henry VI 1\	henry VI 1.txt	142,164
X:\text\shakes\plays.zip\plays\historical\henry VI 2\	henry VI 2.txt	159,602
X:\text\shakes\plays.zip\plays\historical\henry VI 3\	henry VI 3.txt	155,360
X:\text\shakes\plays.zip\plays\historical\henry VIII\	henry VIII.txt	153,653
X:\text\shakes\plays.zip\plays\historical\john\	john.txt	128,089
X:\text\shakes\plays.zip\plays\historical\richard II\	richard II.txt	144,640
X:\text\shakes\plays.zip\plays\historical\richard III\	richard III.txt	196,208

See also : Step-by-step online example, Viewing source texts 116, Finding source texts 430.

5.6.3 favourite texts

save favourites

Used to save your current selection of texts. Useful if it's complex, e.g. involving several different folders. Essential if you've attached a date to your text files.

Saves a list of text files whose status is either unknown or known to meet your requirements when selecting files by their contents 137, ignoring any which do not.

get favourites

Used to read a previously-saved selection from disk.

By default the file name will be the name of the tool you're choosing texts for plus **`recent_chosen_text_files.dat`**, in your main WordSmith folder.

You may use a plain text file for loading (📂) a set of choices you have edited using Notepad, but note that each file needed must be fully specified: wildcards are not used and a full drive:\folder path is needed. You may date the text file if you like by appending to the file-name a <tab> character followed by the date (any date after 1000BC) in the format yyyy/mm/dd e.g.

```
c:\text\socrates.txt    -399/07/01
c:\text\hamlet.txt      1600/07/01
c:\text\second world war.txt  1943/05/22
```

See also: Choosing Texts 44, file dates 48

5.7 choosing files from standard dialogue box

The dialogue box here is very similar to the one used for choosing text files 44; it also allows you to choose from a zip file 453.

You can use Viewer & Aligner 379 to examine a file: this makes no sense in the case of a word list, key word list, or concordance, but may be useful if you need to examine a related text file, e.g. a readme.txt in the same zip file as your concordance or word lists.

To choose more than one file, hold the Control key down as you click with your mouse, to select as many separate files as you want. Or hold down the Shift key to select a whole range of them.

5.8 class or session instructions

When WordSmith is run in a training session, you may want to make certain instructions available to your trainees.

To do this, all you need to do is ensure there is a file called **`teacher.rtf`** in your main **`\wsmith6`** folder where the WordSmith programs are or in the "instructions folder" explained under site licence defaults 23. If one is found, it will be shown automatically when WordSmith starts up. To stop it being shown, just rename it! You edit the file using any Rich Text Format word processor, such as MS Word™, saving as an **`.rtf`** file.

See also: Site licence defaults 23

5.9 colour categories

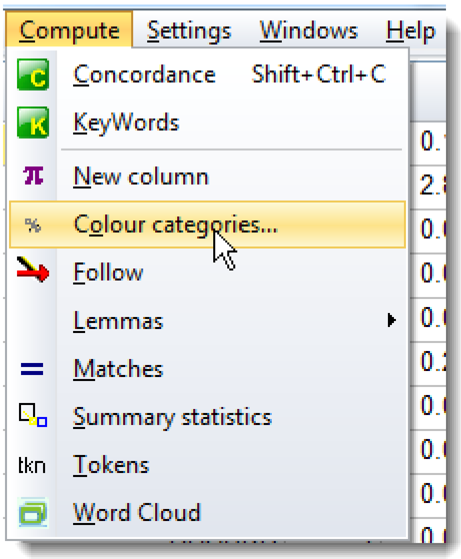
The point of it ...

With a concordance or word list on your screen it can be hard for example to know how many of the thousands of entries met certain criteria. For example which ones derived from only a few texts? Which ones ended in **`-NESS`**? How many of the concordance lines came both from **`mytext.txt`** and from the first 40 words in the sentence, and which ones are they?

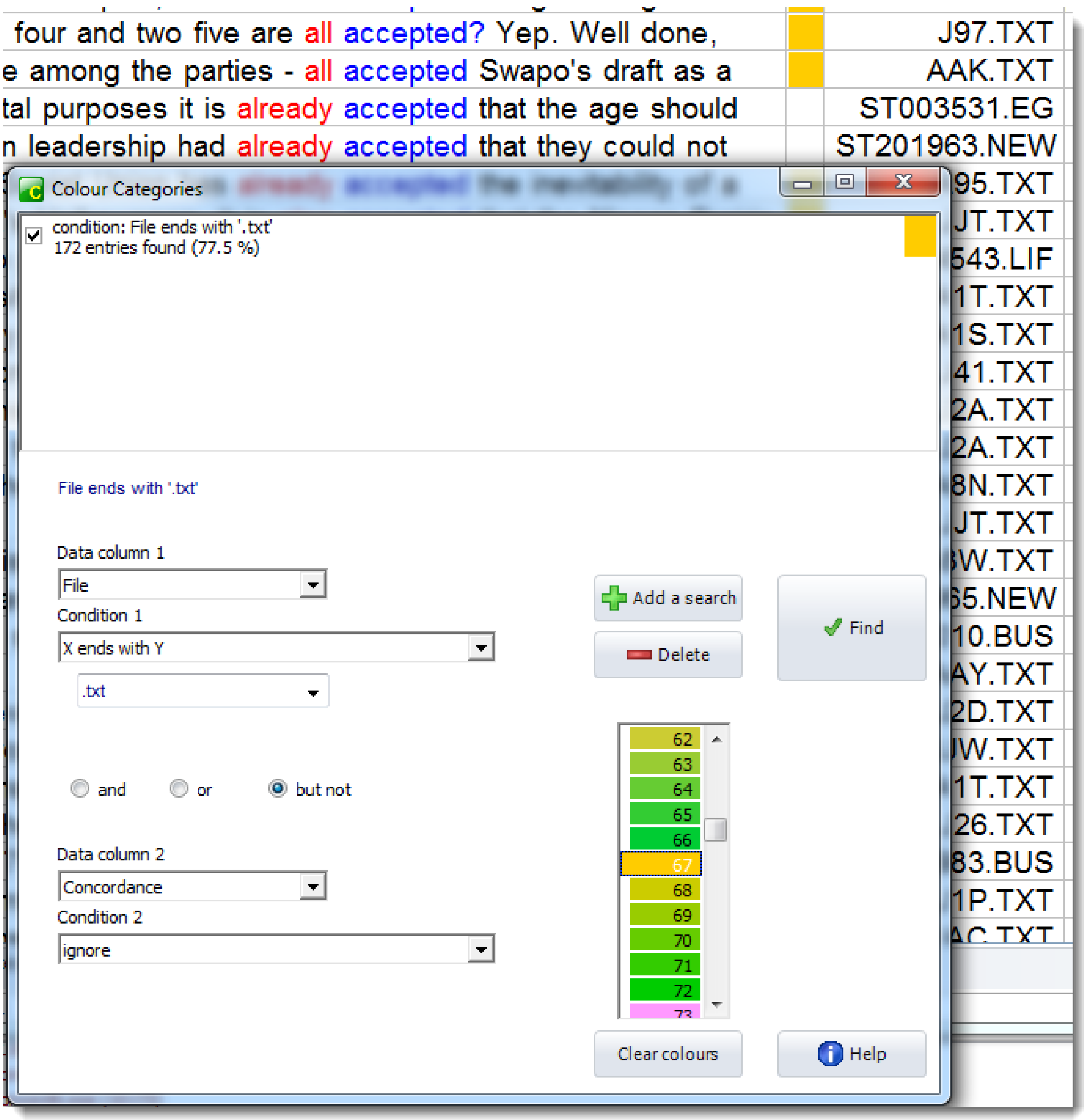
The idea is to let you re-sort 315 your existing data by your own criteria. (Since last millennium WordSmith lists have been sortable by standard criteria, and there has long been a Set 168 column for your own classification, but this feature makes it possible to have multiple and complex sorts.)

How to do it

The menu option *Compute* | *Colour categories* will be found if the data have a Set column.



The menu option brings up a window where you specify your search criteria. Here is an example:



four and two five are all accepted? Yep. Well done, J97.TXT
e among the parties - all accepted Swapo's draft as a AAK.TXT
tal purposes it is already accepted that the age should ST003531.EG
n leadership had already accepted that they could not ST201963.NEW

Colour Categories

☑ condition: File ends with '.txt'
172 entries found (77.5 %)

File ends with '.txt'

Data column 1: File
Condition 1: X ends with Y
.txt

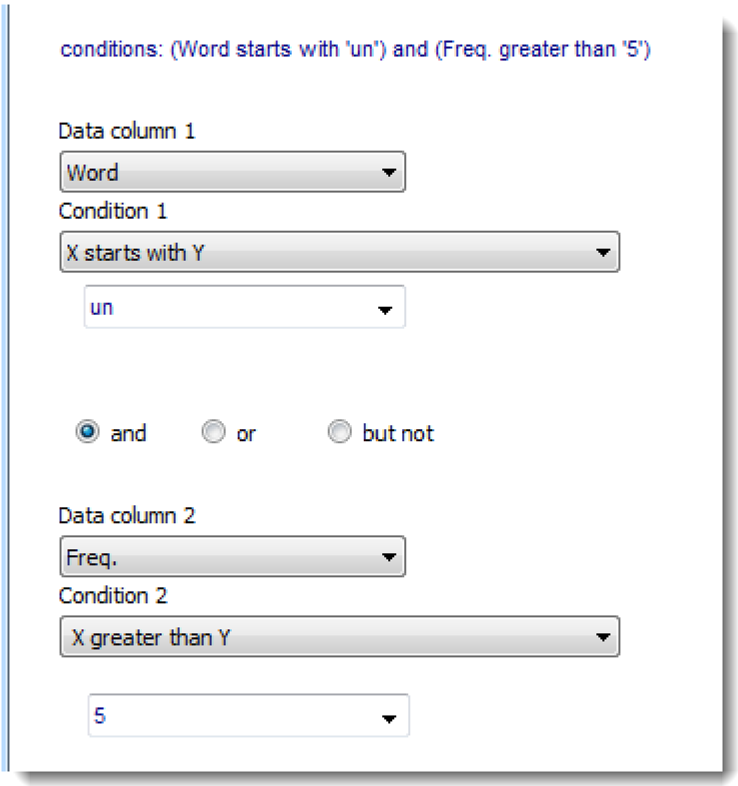
○ and ○ or ● but not

Data column 2: Concordance
Condition 2: ignore

Complete the form by choosing a data column (above the user chose the File column) and a condition (here *X ends with Y* and *.txt* below which will mean 'search the file column seeking any where the File ends in .txt'). Then choose a colour (here colour 67 was chosen) and then press *Add a search*. Finally, press *Find*.

As you can just see, the Set column in the concordance has some items coloured.

A more complex example:



conditions: (Word starts with 'un') and (Freq. greater than '5')

Data column 1: Word

Condition 1: X starts with Y

un

● and ○ or ○ but not

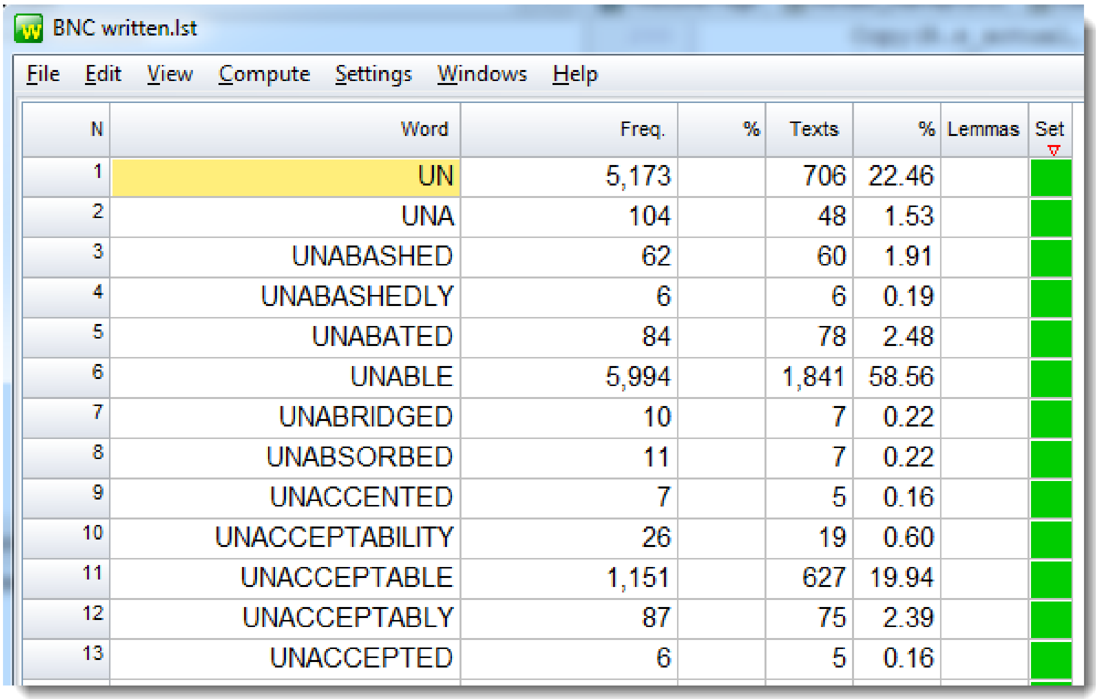
Data column 2: Freq.

Condition 2: X greater than Y

5

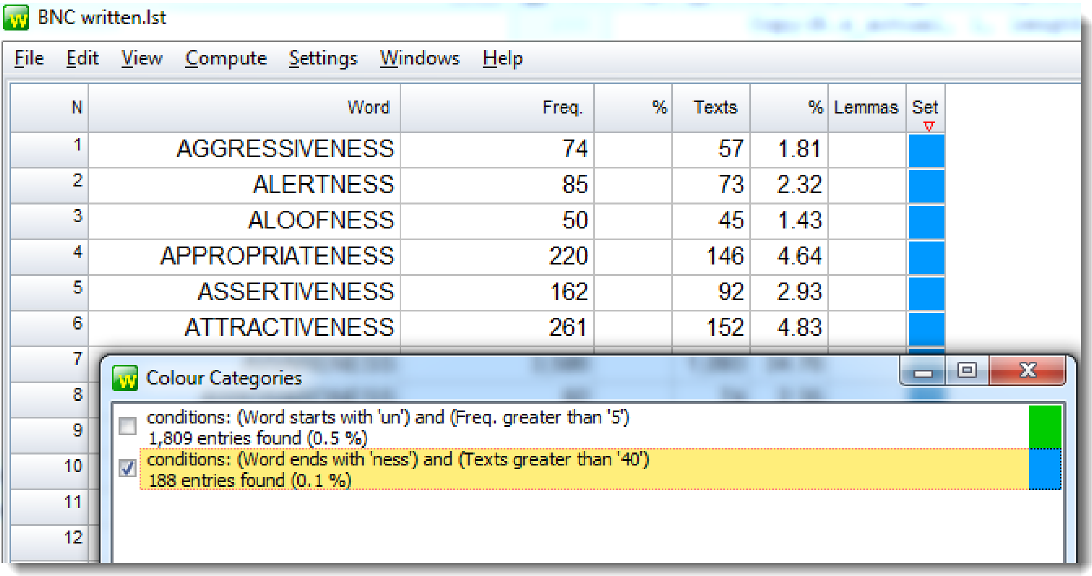
where the user wants to process the Word column of data, looking for a condition where the word starts with **UN** ***and*** occurs at least 5 times. For any word which meets this condition, the Set column will show the colour selected.

When you have specified the criteria, press the *Find* button.



N	Word	Freq.	%	Texts	%	Lemmas	Set
1	UN	5,173		706	22.46		🟩
2	UNA	104		48	1.53		🟩
3	UNABASHED	62		60	1.91		🟩
4	UNABASHEDLY	6		6	0.19		🟩
5	UNABATED	84		78	2.48		🟩
6	UNABLE	5,994		1,841	58.56		🟩
7	UNABRIDGED	10		7	0.22		🟩
8	UNABSORBED	11		7	0.22		🟩
9	UNACCENTED	7		5	0.16		🟩
10	UNACCEPTABILITY	26		19	0.60		🟩
11	UNACCEPTABLE	1,151		627	19.94		🟩
12	UNACCEPTABLY	87		75	2.39		🟩
13	UNACCEPTED	6		5	0.16		🟩

The top of the Colour Categories window shows the percentage results. In the example below, the user has decided to omit their first search and to carry out another on the same word-list which found 188 words ending in **NESS** which were present in more than 40 BNC texts.



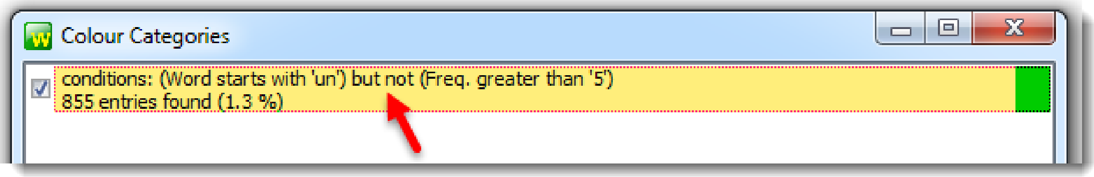
N	Word	Freq.	%	Texts	%	Lemmas	Set
1	AGGRESSIVENESS	74		57	1.81		🟦
2	ALERTNESS	85		73	2.32		🟦
3	ALOOFNESS	50		45	1.43		🟦
4	APPROPRIATENESS	220		146	4.64		🟦
5	ASSERTIVENESS	162		92	2.93		🟦
6	ATTRACTIVENESS	261		152	4.83		🟦
7							
8							
9							
10							
11							
12							

□ conditions: (Word starts with 'un') and (Freq. greater than '5')
1,809 entries found (0.5 %) 🟩

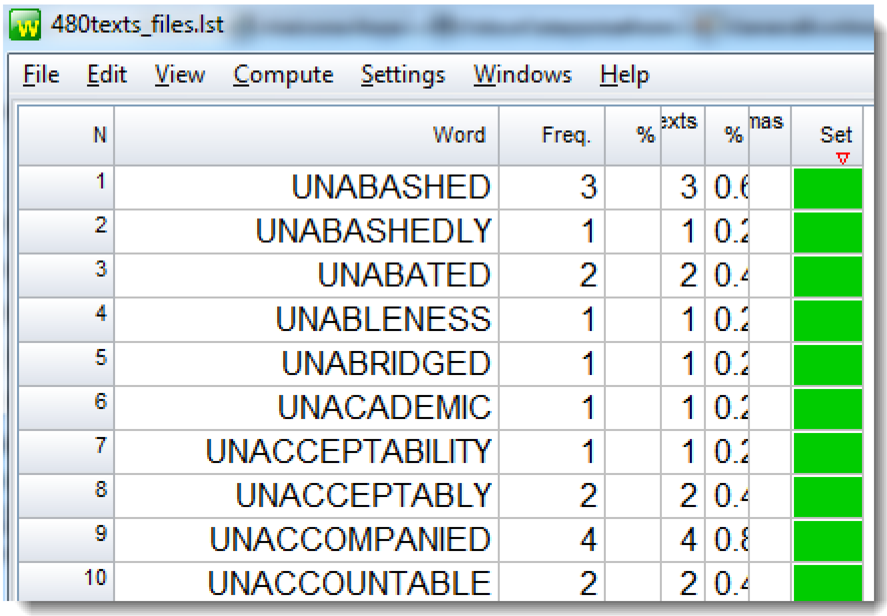
☑ conditions: (Word ends with 'ness') and (Texts greater than '40')
188 entries found (0.1 %) 🟦

But not

This option lets you have a negative condition.



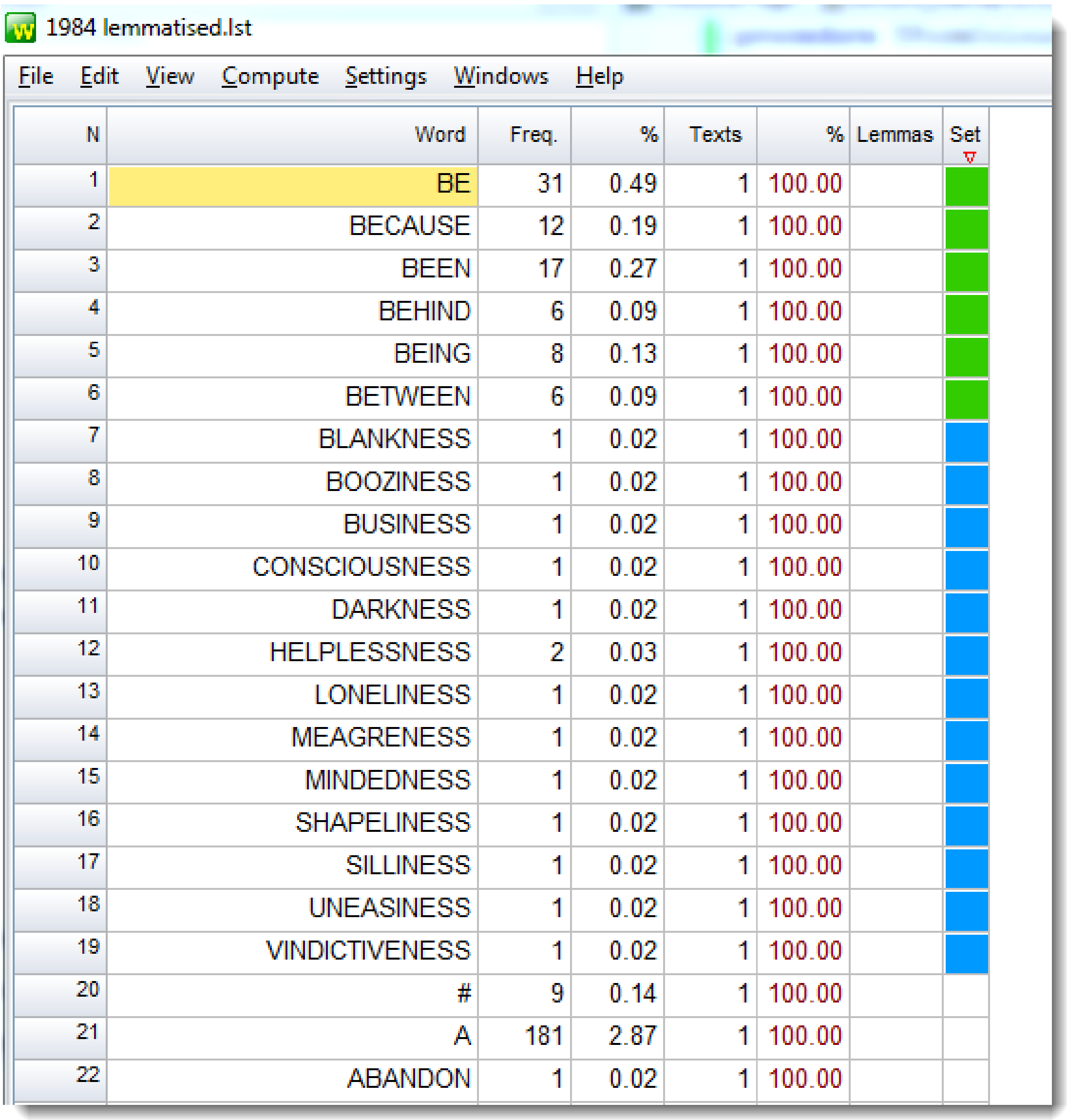
☑ conditions: (Word starts with 'un') but not (Freq. greater than '5')
855 entries found (1.3 %) 🟩



N	Word	Freq.	%	exts	%	mas	Set
1	UNABASHED	3		3	0.6		🟩
2	UNABASHEDLY	1		1	0.2		🟩
3	UNABATED	2		2	0.4		🟩
4	UNABLENESS	1		1	0.2		🟩
5	UNABRIDGED	1		1	0.2		🟩
6	UNACADEMIC	1		1	0.2		🟩
7	UNACCEPTABILITY	1		1	0.2		🟩
8	UNACCEPTABLY	2		2	0.4		🟩
9	UNACCOMPANIED	4		4	0.8		🟩
10	UNACCOUNTABLE	2		2	0.4		🟩

Where are they in the list?

To locate the items which colour categorising has found, simply sort the Set column. (If it's a Freq. list you may have to go to the Alphabetical tab first.) The categorised items float to the top. Here, the 6 words between **BE** and **BF** with frequency above 5 are coloured green at the top of the word list, with the 13 **NESS** items with frequency less than or equal to 5 coloured blue.

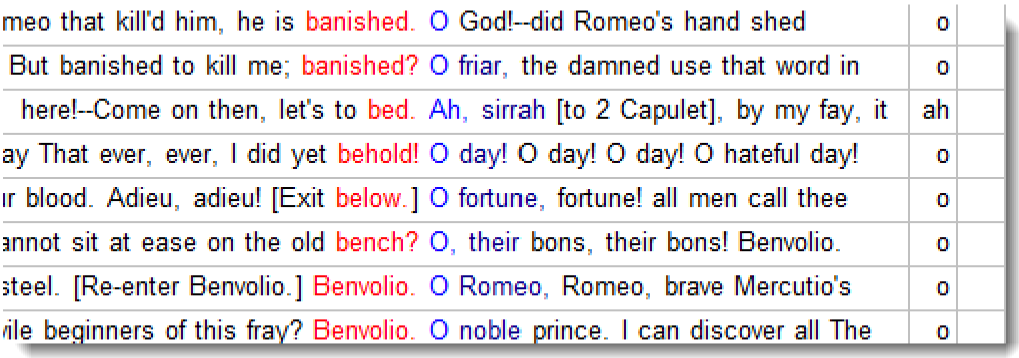


N	Word	Freq.	%	Texts	%	Lemmas	Set
1	BE	31	0.49	1	100.00		🟩
2	BECAUSE	12	0.19	1	100.00		🟩
3	BEEN	17	0.27	1	100.00		🟩
4	BEHIND	6	0.09	1	100.00		🟩
5	BEING	8	0.13	1	100.00		🟩
6	BETWEEN	6	0.09	1	100.00		🟩
7	BLANKNESS	1	0.02	1	100.00		🟦
8	BOOZINESS	1	0.02	1	100.00		🟦
9	BUSINESS	1	0.02	1	100.00		🟦
10	CONSCIOUSNESS	1	0.02	1	100.00		🟦
11	DARKNESS	1	0.02	1	100.00		🟦
12	HELPLESSNESS	2	0.03	1	100.00		🟦
13	LONELINESS	1	0.02	1	100.00		🟦
14	MEAGRENESS	1	0.02	1	100.00		🟦
15	MINDEDNESS	1	0.02	1	100.00		🟦
16	SHAPELINESS	1	0.02	1	100.00		🟦
17	SILLINESS	1	0.02	1	100.00		🟦
18	UNEASINESS	1	0.02	1	100.00		🟦
19	VINDICTIVENESS	1	0.02	1	100.00		🟦
20	#	9	0.14	1	100.00		
21	A	181	2.87	1	100.00		
22	ABANDON	1	0.02	1	100.00		

Once sorted, the data can be saved as before.

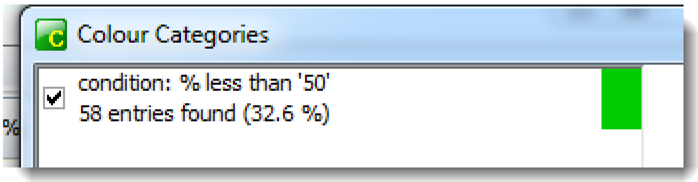
What if I already have a Set classification?

Here is a concordance where the exclamation O or Ah had already been identified and marked in the Set column.



meo that kill'd him, he is banished. O God!--did Romeo's hand shed	o
But banished to kill me; banished? O friar, the damned use that word in	o
here!--Come on then, let's to bed. Ah, sirrah [to 2 Capulet], by my fay, it	ah
ay That ever, ever, I did yet behold! O day! O day! O day! O hateful day!	o
ır blood. Adieu, adieu! [Exit below.] O fortune, fortune! all men call thee	o
annot sit at ease on the old bench? O, their bons, their bons! Benvolio.	o
steel. [Re-enter Benvolio.] Benvolio. O Romeo, Romeo, brave Mercutio's	o
vile beginners of this fray? Benvolio. O noble prince. I can discover all The	o

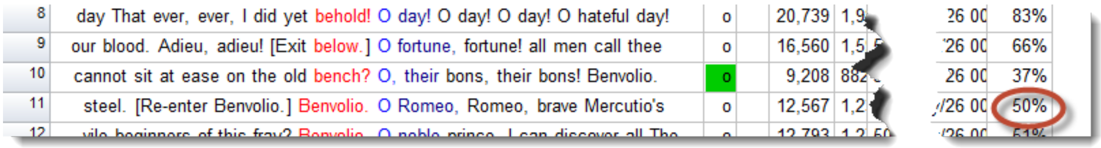
As the Set column is already in use, classifying further by colour will take second priority to the existing forms typed in. So in this case:



Colour Categories

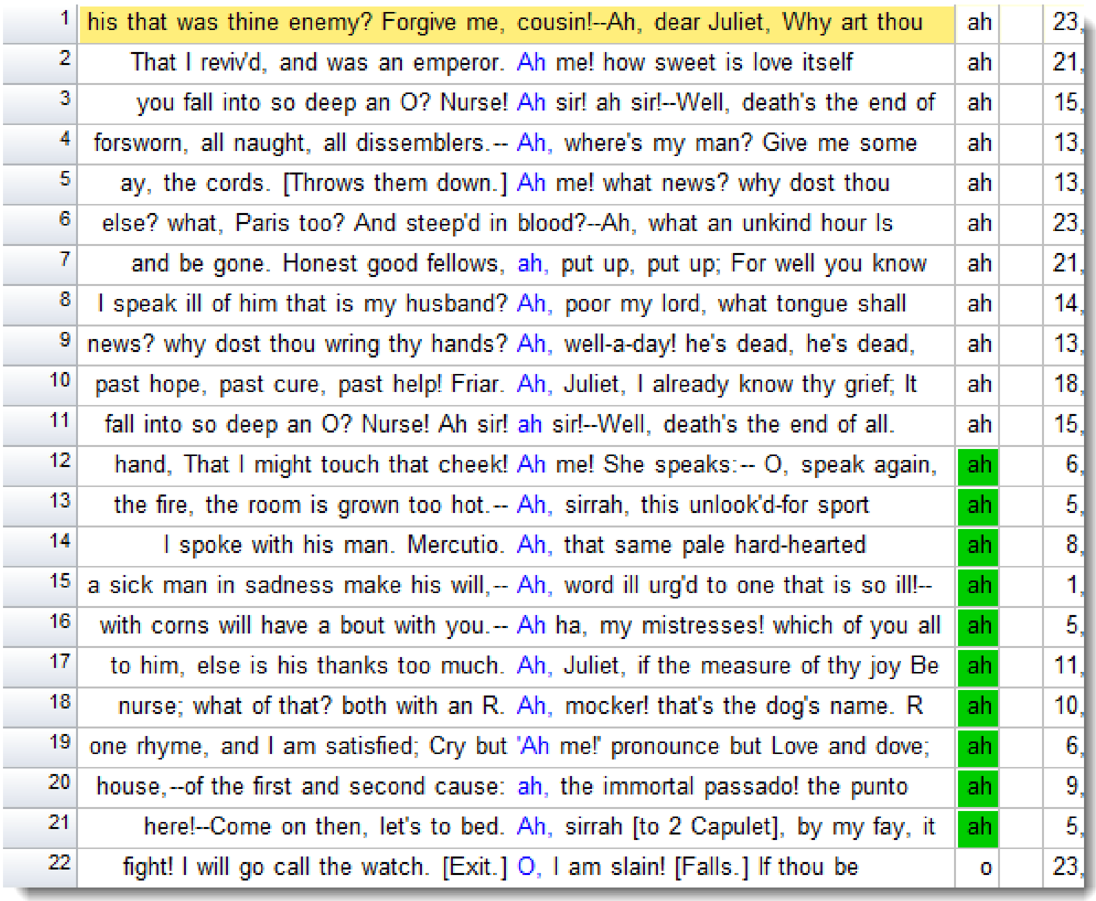
☑ condition: % less than '50'
58 entries found (32.6 %) 🟩

where 58 cases were found where the exclamations came in the first 49% of the text, we see that line 10 goes green (11 did not go green because the criterion was less than 50 and it had exactly 50%)



8	day That ever, ever, I did yet behold! O day! O day! O day! O hateful day!	o	20,739	1,9	26 00	83%
9	our blood. Adieu, adieu! [Exit below.] O fortune, fortune! all men call thee	o	16,560	1,5	26 00	66%
10	cannot sit at ease on the old bench? O, their bons, their bons! Benvolio.	o	9,208	882	26 00	37%
11	steel. [Re-enter Benvolio.] Benvolio. O Romeo, Romeo, brave Mercutio's	o	12,567	1,2	/26 00	50%
12	vile beginners of this fray? Benvolio. O noble prince. I can discover all The	o	12,793	1,2 50	/26 00	51%

but clicking the Set column gives priority to the exclamation typed in.



1	his that was thine enemy? Forgive me, cousin!--Ah, dear Juliet, Why art thou	ah	23,
2	That I reviv'd, and was an emperor. Ah me! how sweet is love itself	ah	21,
3	you fall into so deep an O? Nurse! Ah sir! ah sir!--Well, death's the end of	ah	15,
4	forsworn, all naught, all dissemblers.-- Ah, where's my man? Give me some	ah	13,
5	ay, the cords. [Throws them down.] Ah me! what news? why dost thou	ah	13,
6	else? what, Paris too? And steep'd in blood?--Ah, what an unkind hour Is	ah	23,
7	and be gone. Honest good fellows, ah, put up, put up; For well you know	ah	21,
8	I speak ill of him that is my husband? Ah, poor my lord, what tongue shall	ah	14,
9	news? why dost thou wring thy hands? Ah, well-a-day! he's dead, he's dead,	ah	13,
10	past hope, past cure, past help! Friar. Ah, Juliet, I already know thy grief; It	ah	18,
11	fall into so deep an O? Nurse! Ah sir! ah sir!--Well, death's the end of all.	ah	15,
12	hand, That I might touch that cheek! Ah me! She speaks:-- O, speak again,	ah	6,
13	the fire, the room is grown too hot.-- Ah, sirrah, this unlook'd-for sport	ah	5,
14	I spoke with his man. Mercutio. Ah, that same pale hard-hearted	ah	8,
15	a sick man in sadness make his will,-- Ah, word ill urg'd to one that is so ill!--	ah	1,
16	with corns will have a bout with you.-- Ah ha, my mistresses! which of you all	ah	5,
17	to him, else is his thanks too much. Ah, Juliet, if the measure of thy joy Be	ah	11,
18	nurse; what of that? both with an R. Ah, mocker! that's the dog's name. R	ah	10,
19	one rhyme, and I am satisfied; Cry but 'Ah me!' pronounce but Love and dove;	ah	6,
20	house,--of the first and second cause: ah, the immortal passado! the punto	ah	9,
21	here!--Come on then, let's to bed. Ah, sirrah [to 2 Capulet], by my fay, it	ah	5,
22	fight! I will go call the watch. [Exit.] O, I am slain! [Falls.] If thou be	o	23,

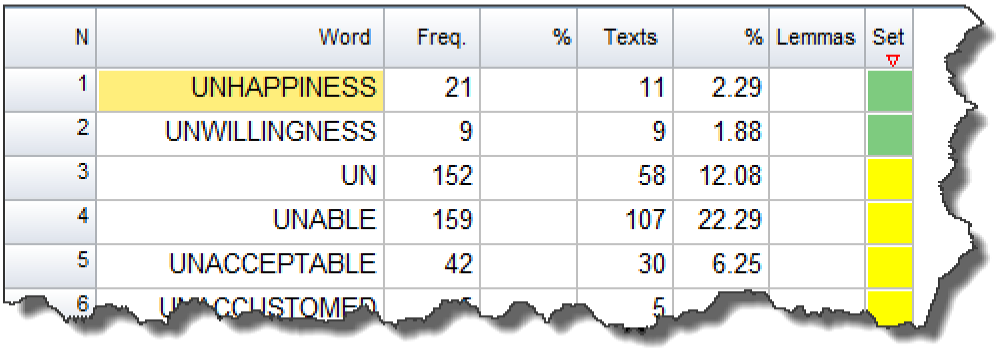
In this case the **O**s follow the **Ah**s, and the coloured **Ah**s follow the uncoloured ones.

Removing the colours?

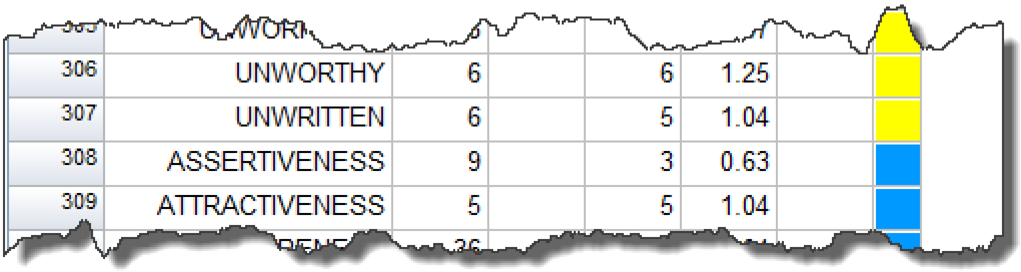
Use the Clear colours button.

What if more than one condition is met?

If you colour words ending **NESS** blue, and also colour words starting **UN** yellow, any word meeting both conditions will get a mixture of the two colours as shown here:



N	Word	Freq.	%	Texts	%	Lemmas	Set
1	UNHAPPINESS	21		11	2.29		🟩
2	UNWILLINGNESS	9		9	1.88		🟩
3	UN	152		58	12.08		🟨
4	UNABLE	159		107	22.29		🟨
5	UNACCEPTABLE	42		30	6.25		🟨
6	UNACCUSTOMED			5			🟨



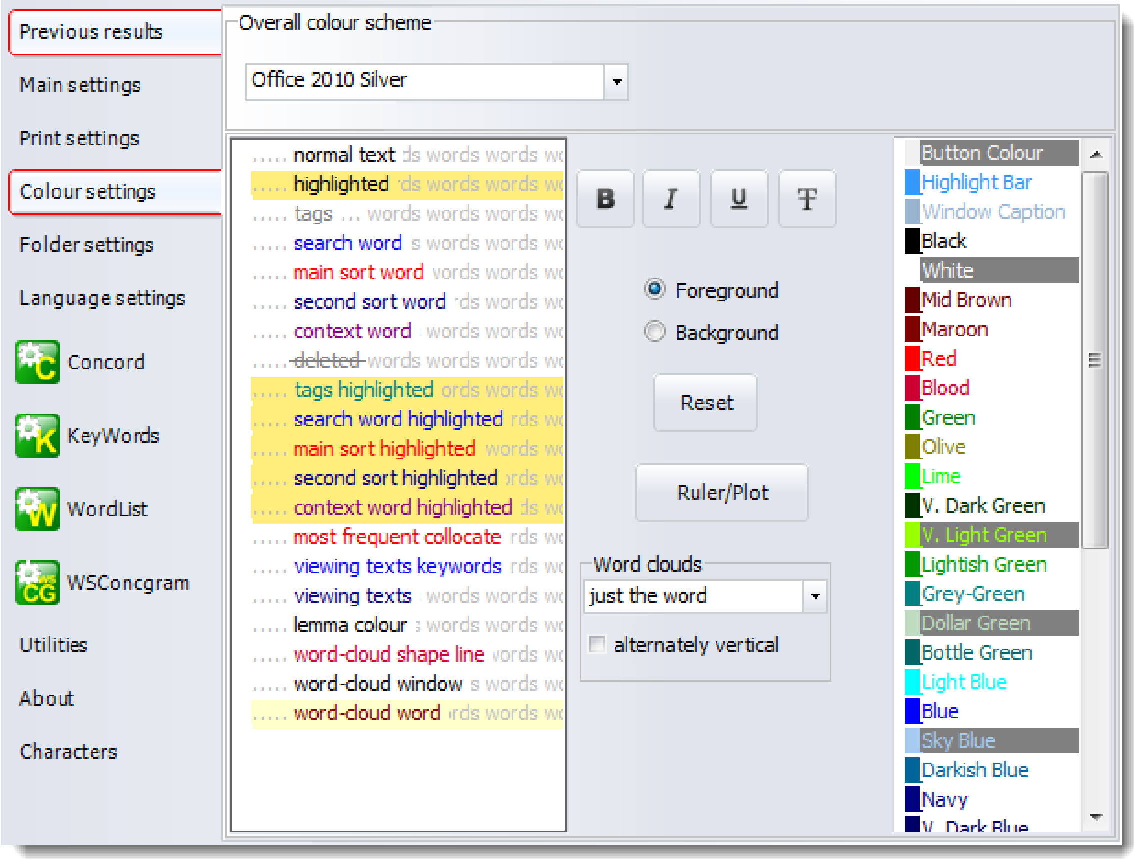
306	UNWORTHY	6		6	1.25		🟨
307	UNWRITTEN	6		5	1.04		🟨
308	ASSERTIVENESS	9		3	0.63		🟦
309	ATTRACTIVENESS	5		5	1.04		🟦

See also : setting categories by typing 168, colour categories for concordances 171

5.10 colours

Found in main Settings menu in all Tools and Main Settings in the Controller 4. Enables you to choose your default colours for all the Tools. Available colours can be set for

plain text	this is the default colour
highlighted text	as above when selected
tags 198	mark-up
search word 159	concordance search word; words in (key) word lists
main sort word 315	indicates first sort preference; used for % data in (key) word lists
second sort word	indicates first tie-breaker sort colour
context word	context word
deleted words	any line of deleted data
not numbered line	any line which has not been user-sorted
search word highlighted	concordance search word when selected
main sort word highlighted	first sort when selected
second sort word highlighted	first tie-breaker sort when selected
context word highlighted	context word when selected
most frequent collocate	most frequent collocate or detailed consistency word, p value 235 in keywords
viewing texts	in the text viewer 11
lemma colour	colour of lemmas shown in lemma window
word-cloud shape	see word clouds section below
word-cloud window	
word-cloud word	



Overall colour scheme: Office 2010 Silver

● Foreground ○ Background

Word clouds: just the word

□ alternately vertical

Overall colour scheme

This allows a range of colour scheme choices, which will affect the colours of all WordSmith windows.

List colours

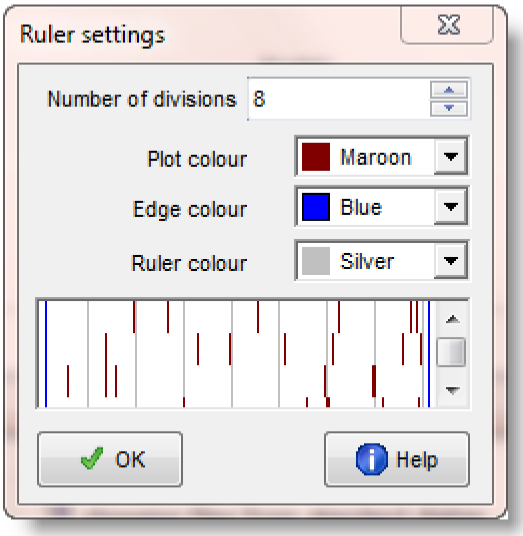
To alter colours, first click on the wording you wish to change (you'll see a difference in the left margin: here search word has been chosen), then click on a colour in the colour box. The *Foreground* and *Background* radio buttons determine whether you're changing foreground or background colours. You can press the Reset button if you want to revert to standard defaults.

The same colours, or equivalent shades of grey, will appear in printouts, or you can set the printer 80 to black and white, in which case any column not using "plain text" colour will appear in italics (or bold or underlined if you have already set the column to italics).

The *Reset* button lets you restore colours to factory defaults.

Ruler

This opens another dialogue window, in which you can set colours and plot divisions for the ruler:



Ruler settings

Number of divisions 8

Plot colour Maroon

Edge colour Blue

Ruler colour Silver

Word Clouds

These settings allow to to choose how each word will be displayed, e.g. within rectangles or circles. The colours of the words and the word cloud window are set in the List colours section above.

See also: Column Layout 87 for changing the individual colours of each column of data, Colour Categories 51, Word Clouds 128.

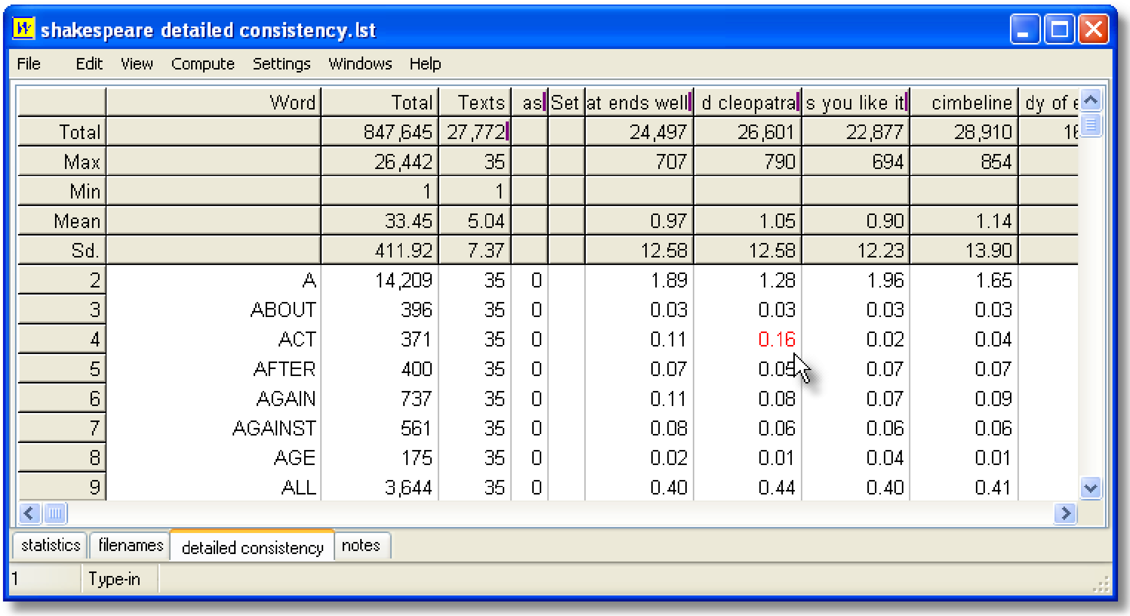
5.11 column totals

The point of it...

This function allows you to see a total and basic statistics on each column of data, if the data are numerical.

How to do it

With a word-list, concordance or key-words list visible, choose the menu item *View* | *Column Totals* to switch column totals on or off.



	Word	Total	Texts	as	Set	at ends well	d cleopatra	s you like it	cimbeline	dy of e
Total		847,645	27,772			24,497	26,601	22,877	28,910	16
Max		26,442	35			707	790	694	854	
Min		1	1							
Mean		33.45	5.04			0.97	1.05	0.90	1.14	
Sd.		411.92	7.37			12.58	12.58	12.23	13.90	
2	A	14,209	35	0		1.89	1.28	1.96	1.65	
3	ABOUT	396	35	0		0.03	0.03	0.03	0.03	
4	ACT	371	35	0		0.11	0.16	0.02	0.04	
5	AFTER	400	35	0		0.07	0.05	0.07	0.07	
6	AGAIN	737	35	0		0.11	0.08	0.07	0.09	
7	AGAINST	561	35	0		0.08	0.06	0.06	0.06	
8	AGE	175	35	0		0.02	0.01	0.04	0.01	
9	ALL	3,644	35	0		0.40	0.44	0.40	0.41	

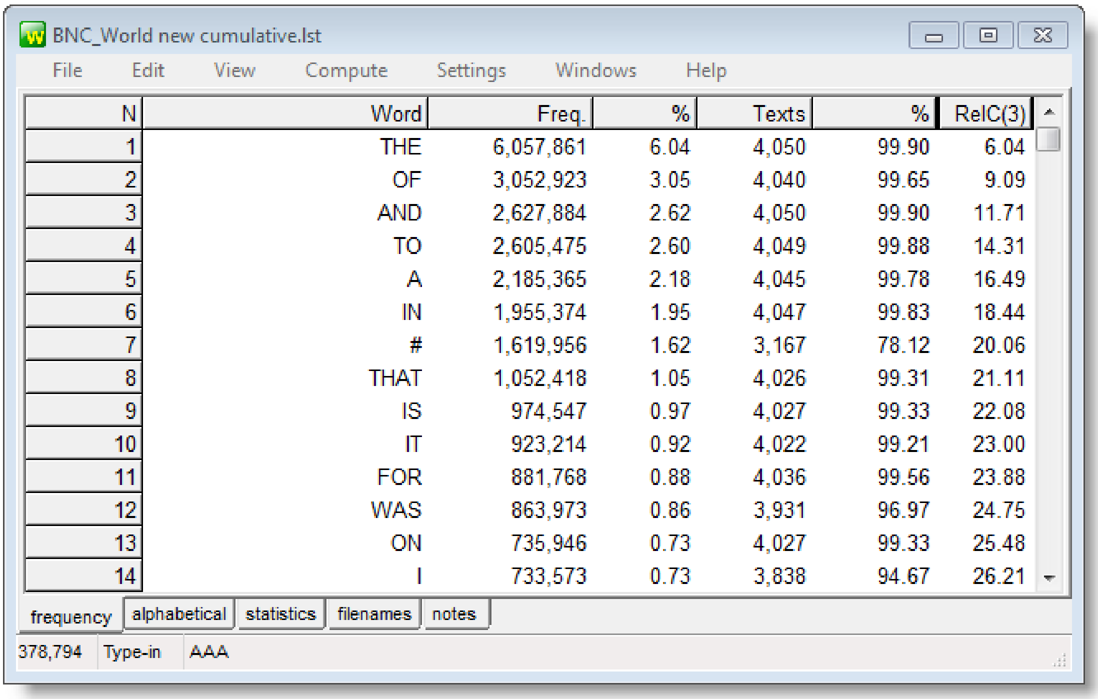
Here we see column totals on a detailed consistency list based on Shakespeare's plays. The list itself is sorted by the Texts column: the top items are found in all 35 of the plays used for the list. In the case of Anthony and Cleopatra, *A* represents 1.28% of the words in that column, that is 1.28% of the words of the play Anthony and Cleopatra. In the case of *ACT* this is the highest percentage in its row (this word is used more in percentage terms in that play than in the others).

See also: save as Excel 102

5.12 compute new column of data

The point of it...

This function brings up a calculator, where you can choose functions to calculate values which interest you. For example, a word list routinely provides the frequency of each type, and that frequency as a percentage of the overall text tokens. You might want to insert a further column showing the frequency as a percentage of the number of word types, or a column showing the frequency as a percentage of the number of text files from which the word list was created.

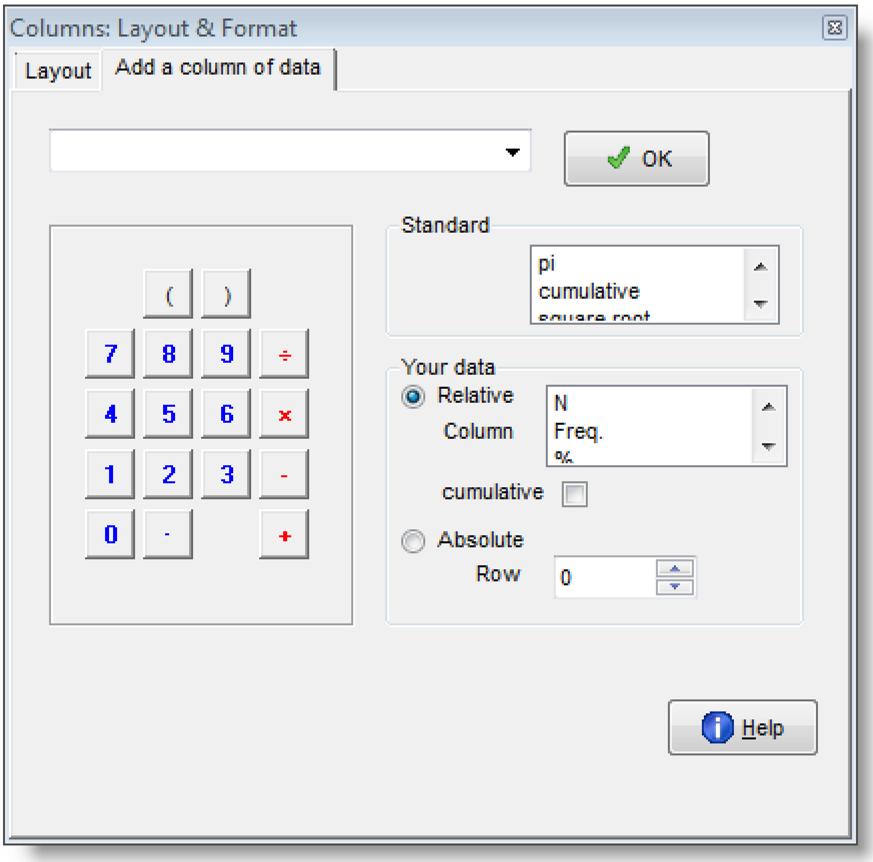


N	Word	Freq.	%	Texts	%	RelC(3)
1	THE	6,057,861	6.04	4,050	99.90	6.04
2	OF	3,052,923	3.05	4,040	99.65	9.09
3	AND	2,627,884	2.62	4,050	99.90	11.71
4	TO	2,605,475	2.60	4,049	99.88	14.31
5	A	2,185,365	2.18	4,045	99.78	16.49
6	IN	1,955,374	1.95	4,047	99.83	18.44
7	#	1,619,956	1.62	3,167	78.12	20.06
8	THAT	1,052,418	1.05	4,026	99.31	21.11
9	IS	974,547	0.97	4,027	99.33	22.08
10	IT	923,214	0.92	4,022	99.21	23.00
11	FOR	881,768	0.88	4,036	99.56	23.88
12	WAS	863,973	0.86	3,931	96.97	24.75
13	ON	735,946	0.73	4,027	99.33	25.48
14	I	733,573	0.73	3,838	94.67	26.21

This word-list has a column which computes the cumulative scores (running total of the % column).

How to do it

Just press *Compute* | *New Column* and create your own formula. You'll see standard calculator buttons with the numbers 0 to 9, decimal point, brackets, 4 basic functions. To the right there's a list of standard mathematical functions to use (pi, square root etc.): to access these, double-click on them. Below that you will see access to your own data in the current list, listing any number-based column-headings. You can drag or double-click them too.



Absolute and Relative

Your own data can be accessed in two ways. A relative access (the default) means that as in a spreadsheet you want the new column to access data from another column but in the same row. Absolute access means accessing a fixed column and row.

Examples

you type	**Result -- for each row in your data, the new column will contain:**
Rel(2) ÷ 5	the data from column 2 of the same row, divided by 5
RelC(2)	the data from column 2 of the same row, added to a running total
Rel(3) + (Rel(2) ÷ 5)	the data from column 2 of the same row, divided by 5, added t the data from column 3 of the same row
Abs(2;1) ÷ 5	the data from column 2 of row 1, divided by 5. (This example is just to illustrate; it would be silly as it would give the exact same result in every row.)
Rel(2) ÷ Abs(2;1) × 100	the data from column 2 of the same row divided by column 2 c row 1 and multiplied by 100. This would give column 3 as a percentage of the top result in column 2. For the first row it'd give 100%, but as the frequencies declined so would their

percentage of the most frequent item.

You can format (or even delete) any variables computed in this way: see layout 87.

See also: count data frequencies 66, column totals 62, colour categories 51

5.13 copy your results

The quickest and easiest method of copying your data e.g. into your word processor is to select with the cursor arrows and then press Ctrl+C. This puts it into the clipboard 422 (click to see examples showing how to copy into Word etc.).

If you choose *File* | *Save As* you get various choices:

saving as a text file or XML or spreadsheet 102
save 101 as data (not the same as saving as text: this is saving so you can access your data again another day)

See also: saving 101, printing 97, clipboard 422

5.14 count data frequencies

In various Tools you may wish to further analyse your data. For example with a concordance you may want to know how many of the lines contain a prefix like **`un-`** or how many items in a word-list end in **`-ly`**. To do this, choose *Summary Statistics* in the *Compute* menu.

Load

This allows you to load into the searches window any plain text file which you have prepared previously. For complex searching this can save much typing. An example might be a list of suffixes or prefixes to check against a word list.

Search Column

This lets you choose which column of data to count in. It will default to the last column clicked for your data.

Breakdown by

If activated this lets you break down results, for example by text file. See the example from Concord 215.

Cumulative Column

This adds up values from another column of data. See the example from WordList 304.

See also: distinguishing consequence from consequences 215, frequencies of suffixes in a word list 304, compute new column of data 63.

5.15 custom processing

This feature -- which, like API 416, is not for those without a tame programmer to help -- is found under *Main Settings* | *Advanced*.

The point of it...

I cannot know which criteria you have in processing your texts, other than the criteria already set up (the choice of texts, of search-word, etc.) You might need to do some specialised checks or alteration of data before it enters the **WordSmith** formats. For example, you might need to lemmatise a word according to the special requirements of your language.

This function makes that possible. If for example you have chosen to filter concordances, as **Concord** processes your text files, every time it finds a match for your search-word, it will call your **`.dll`** file. It'll tell your own **`.dll`** what it has found, and give it a chance to alter the result or tell **Concord** to ignore this one.

How to do it...

Choose your **`.dll`** file (it can have any filename you've chosen for it) and check one or more of the options in the Advanced page. You will need to call standard functions and need to know their names and formats. It is up to you to write your own .dll program which can do the job you want. This can be written in any programming language (C++, Java, Pascal, etc.).

An example for lemmatising a word in WordList

The following DLL is supplied with your installation, compiled & ready to run.

Your .dll needs to contain a function with the following specifications

```
function WordlistChangeWord(
  original : pointer;
  language_identifier : DWORD;
  is_Unicode : WordBool) : pointer; stdcall;
```

The language_identifier is a number corresponding to the language you're working with. See List of Locale ID (LCID) Values as Assigned by Microsoft .

So the "original" (sent by WordSmith) can be a PCHAR (7 or 8-bit) or a PWIDECHAR (16-bit Unicode) and the result which your .dll supplies can point to

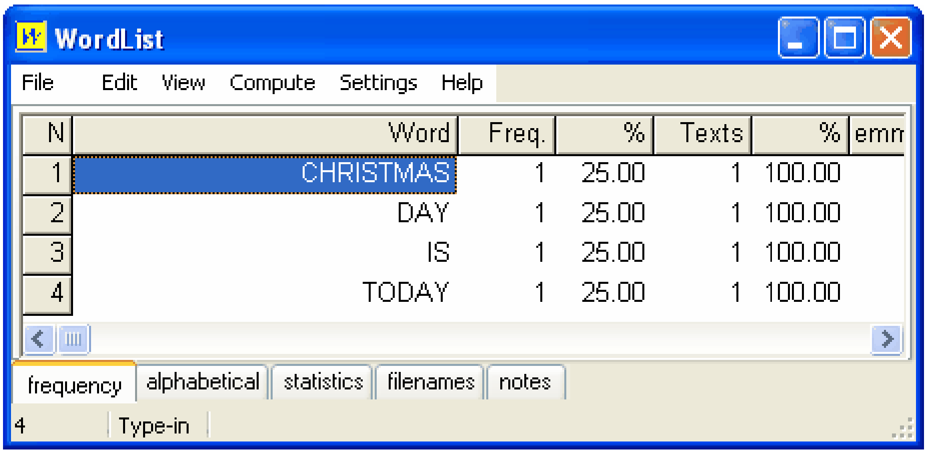
a) nil (if you simply do not want the original word in your list)

b) the same PCHAR/PWIDECHAR if it is not to be changed at all

c) a replacement form

Here's an example where the source text was

`Today is Easter Day.`



N	Word	Freq.	%	Texts	%	emm
1	CHRISTMAS	1	25.00	1	100.00	
2	DAY	1	25.00	1	100.00	
3	IS	1	25.00	1	100.00	
4	TODAY	1	25.00	1	100.00	

Source code

The source code for the .dll in Delphi is this

```
library WS5WordSmithCustomDLL;

uses

  Windows, SysUtils;

{
 This example uses a very straightforward Windows routine for comparing
 strings, CompareStringA and CompareStringW which are in a Windows .dll.

 The function does a case-insensitive comparison because
 NORM_IGNORECASE (=1) is used. If it was replaced by 0, the comparison
 would be case-sensitive.

 In this example, EASTER gets changed to CHRISTMAS.
}

function WordlistChangeWord(
  original : pointer;
  language_identifier : DWORD;
  is_Unicode : WordBool) : pointer; stdcall;
begin
  Result := original;
  if is_Unicode then begin
    if CompareStringW(
      language_identifier,
      NORM_IGNORECASE,
      PWideChar(original), -1,
      PWideChar(widestring('EASTER')), -1) - 2 = 0
    then
      Result := pwidechar(widestring('CHRISTMAS'));
  end else begin
    if CompareStringA(
      language_identifier,
      NORM_IGNORECASE,
      PAnsiChar(original), -1,
      PAnsiChar('EASTER'), -1) - 2 = 0
    then
      Result := pAnsichar('CHRISTMAS');
  end;
end;

function ConcordChangeWord(
  original : pointer;
  language_identifier : DWORD;
  is_Unicode : WordBool) : pointer; stdcall;
begin
  Result := WordlistChangeWord(original,language_identifier,is_unicode);
end;

function KeyWordsChangeWord(
  original : pointer;
  language_identifier : DWORD;
  is_Unicode : WordBool) : pointer; stdcall;
begin
  Result := WordlistChangeWord(original,language_identifier,is_unicode);
end;

{
 This routine exports each concordance line together with
   the filename it was found in
   a number stating how many bytes into the source text file the entry was fou
   its hit position in that text file counted in characters (not bytes) and
   the length of the hit-word
   (so if the search was on HAPP* and the hit was HAPPINESS this would be 9)
 This information is saved in Unicode appended to your results_filename
}

function HandleConcordanceLine
 (source_line : pointer;
  hit_pos_in_characters,
  hit_length : integer;
  byte_position_in_file,
  language_id : DWORD;
  is_Unicode : WordBool;
  source_text_filename,
  results_filename : pwidechar) : pointer; stdcall;

  function extrasA : ansistring;
  begin
    Result := #9+ ansistring(widestring(pwidechar(source_text_filename)))+
              #9+ ansistring(IntToStr(byte_position_in_file))+
              #9+ ansistring(IntToStr(hit_pos_in_characters))+
              #9+ ansistring(IntToStr(hit_length));
  end;

  function extrasW : widestring;
  begin
    Result := #9+ widestring(pwidechar(source_text_filename))+
              #9+ IntToStr(byte_position_in_file)+
              #9+ IntToStr(hit_pos_in_characters)+
              #9+ IntToStr(hit_length);
  end;

const
  bm: char = widechar($FEFF);
var f : File of widechar;
  output_string : widestring;
begin
  Result := source_line;
  if length(results_filename)>0 then
  try
    AssignFile(f,results_filename);
    if FileExists(results_filename) then begin
      Reset(f);
      Seek(f, FileSize(f));
    end else begin
      Rewrite(f);
      Write(f, bm);
    end;
    if is_Unicode then
      output_string := pwidechar(source_line)+extrasW
    else
      output_string := pAnsichar(source_line)+widestring(extrasA);
    if length(output_string) > 0 then
      BlockWrite(f, output_string[1], length(output_string));
    CloseFile(f);
  except
  end;
end;

exports

  ConcordChangeWord,
  KeyWordsChangeWord,
  WordlistChangeWord,
  HandleConcordanceLine;

begin
end.
```

See also : API 416, custom settings 133

5.16 editing

5.16.1 reduce data to n entries

With a very large word-list, concordance etc., you may wish to reduce it randomly (eg. for sampling). This menu option (*Edit* | *Deleting* | *Reduce to N*) allows you to specify how many entries you want to have in the list. If you reduce the data, entries will be randomly zapped 129 until there are only the number you want. The procedure is irreversible. That is, nothing gets altered on disk, but if you change your mind you will have to re-compute or else go back to an earlier saved version.

See also: zapping 129, editing a list of data 72.

5.16.2 reverse-delete

This function simply reverses any deletions in your entries. That is, any entries which were not previously marked as deleted get deleted, and any which were marked as deleted get restored.

See also : Delete or Restore to End 71.

5.16.3 delete or restore to end

These menu options let you mark all entries from the current selected entry downwards as deleted -- or alternatively restore to the end.

See also: Reverse-delete 71, Reduce to N entries 70

5.16.4 delete if

The idea is to be able to delete entries with a search.

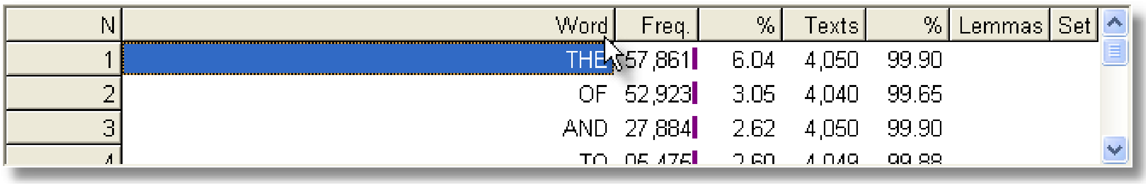
The search operates on the column of data which you have currently selected, so first ensure you click the data in the desired column.

The syntax 159 is as in Concord, so you may need to use asterisks.

If you are searching a concordance line, the search will operate on the whole of the line that Concord knows about, not just the few words you can see.

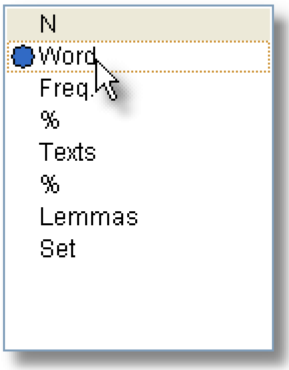
5.16.5 editing column headings

By default, a word-list will have column headings like these:

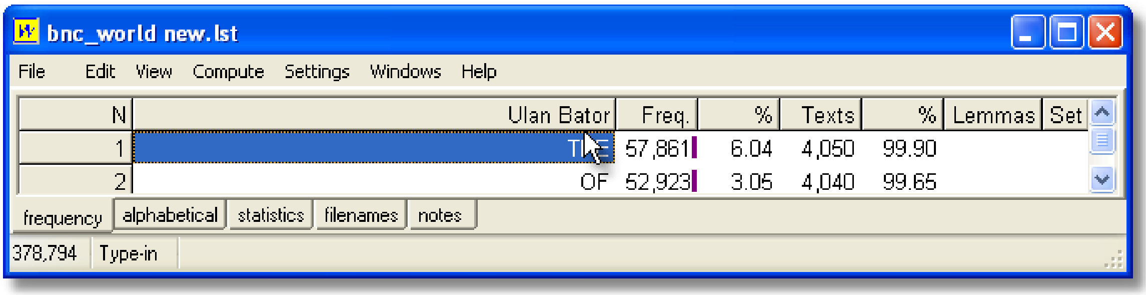


N	Word	Freq.	%	Texts	%	Lemmas	Set
1	THE	57,861	6.04	4,050	99.90		
2	OF	52,923	3.05	4,040	99.65		
3	AND	27,884	2.62	4,050	99.90		
4	TO	[illegible]	2.60	4,049	99.88		

If you choose *View* | *Layout,* you get to see the various headings:



and if you double-click any of these you may edit it to change the column header as in this (absurd) example:



N	Ulan Bator	Freq.	%	Texts	%	Lemmas	Set
1	THE	57,861	6.04	4,050	99.90		
2	OF	52,923	3.05	4,040	99.65		

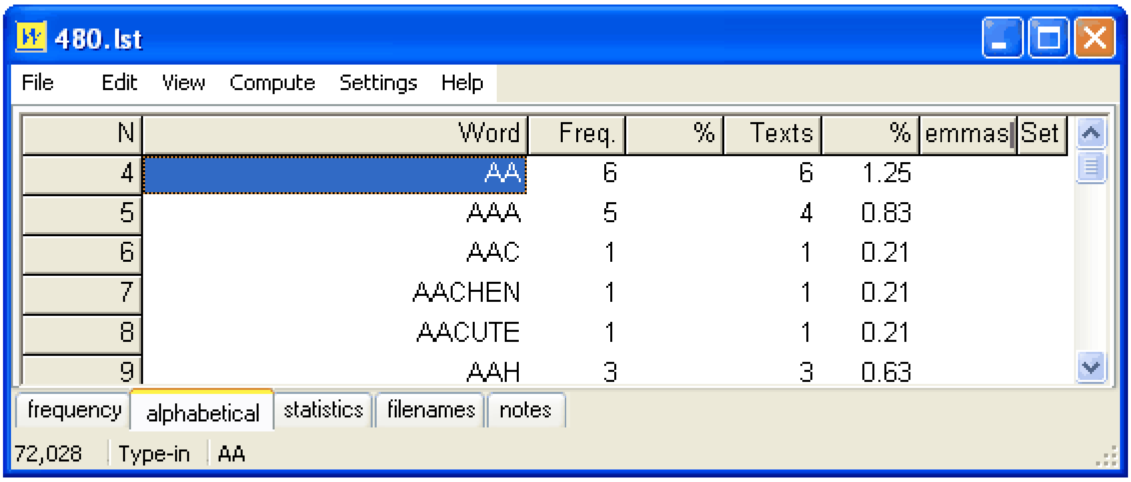
If you now save your word-list, the new column heading gets saved along with the data. Other new word-lists, though, will have the default WordSmith headings.

If you want *all future* word-lists to have the same headings, you should press the Save button in the layout window 87.

(If you had been silly enough to call the word column "Ulan Bator" and to have saved this for all subsequent word-lists, you could remedy the problem by deleting **`Documents\wsmith6\wordlist list customised.dat`**.)

5.16.6 editing a list of data

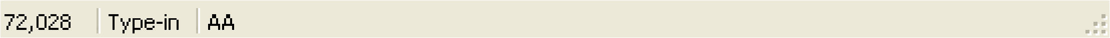
With a word list on screen, you might see something like this.



N	Word	Freq.	%	Texts	%	emmas	Set
4	AA	6		6	1.25		
5	AAA	5		4	0.83		
6	AAC	1		1	0.21		
7	AACHEN	1		1	0.21		
8	AACUTE	1		1	0.21		
9	AAH	3		3	0.63		

72,028 | Type-in | AA

In the status bar at the bottom,

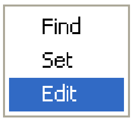


72,028 | Type-in | AA

the number in the first cell is the number of words in the current word list and **AA** in the third cell is the word selected.

At the moment, when the user types anything, WordList will try to find what is typed in the list.

If you right-click the second cell you will see



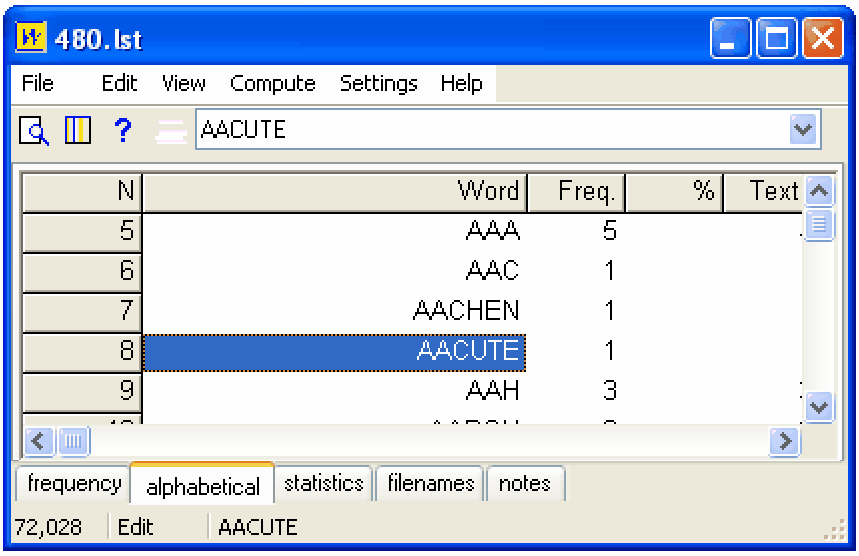
and can change the options for this list to *Set* (to classify your words, eg. as adjectives v. nouns) or *Edit*, to alter them. Note that some of the data is calculated using other data and therefore cannot be edited. For example, frequency percentage data is based on a word's frequency and the total number of running words. You can edit the word frequency but not the word frequency percentage.

Choose *Edit.*

Now, in the column which you want to edit, press any letter.

This will show the toolbar (if it wasn't visible before) so you can alter the form of the word or its frequency. If you spell the word so that it matches another existing word in the list, the list will be altered to reflect your changes.

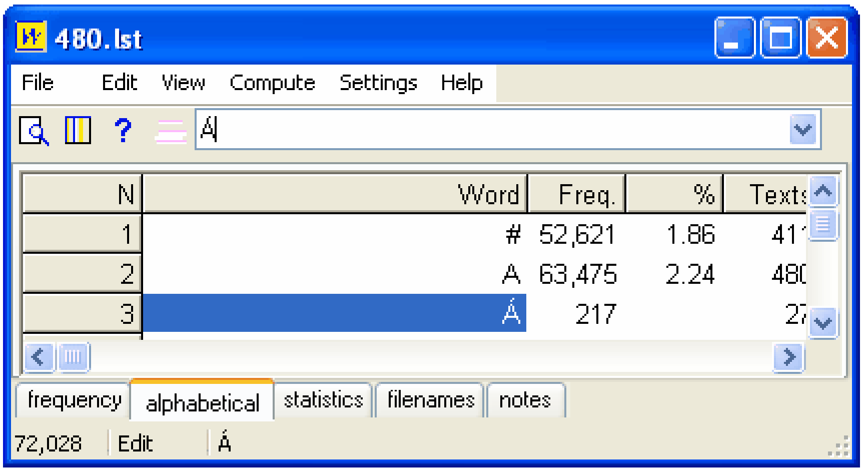
In this case we want to correct **AACUTE**, which should be **Á**.



N	Word	Freq.	%	Text
5	AAA	5		
6	AAC	1		
7	AACHEN	1		
8	AACUTE	1		
9	AAH	3		

72,028 Edit AACUTE

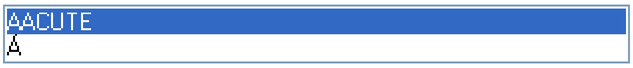
If you now type **Á**, you will immediately see the result in the window:



N	Word	Freq.	%	Texts
1	#	52,621	1.86	41
2	A	63,475	2.24	48
3	Á	217		2

72,028 Edit Á

Clicking the downward arrow at the right of the edit combobox, you will see that the original word is there just in case you decide to retain it.



AACUTE
Á

After editing you may want to re-sort 441 ( ), and if you have changed a word such as **AAAAAGH** to a pre-existing word such as **AAGH**, to join 270 the two entries.

See also: joining entries 270, finding source files 430.

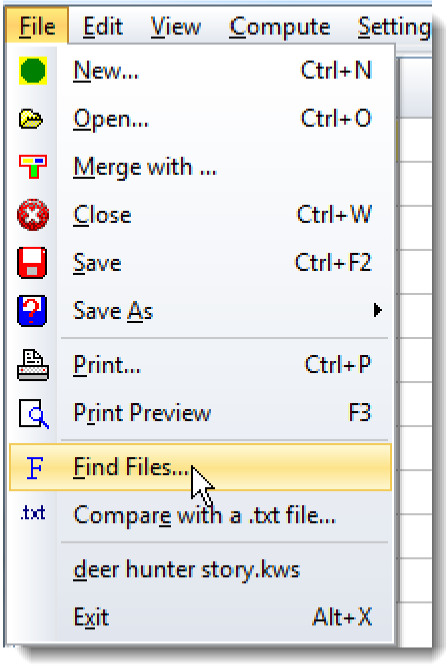
5.17 find relevant files

The point of it...

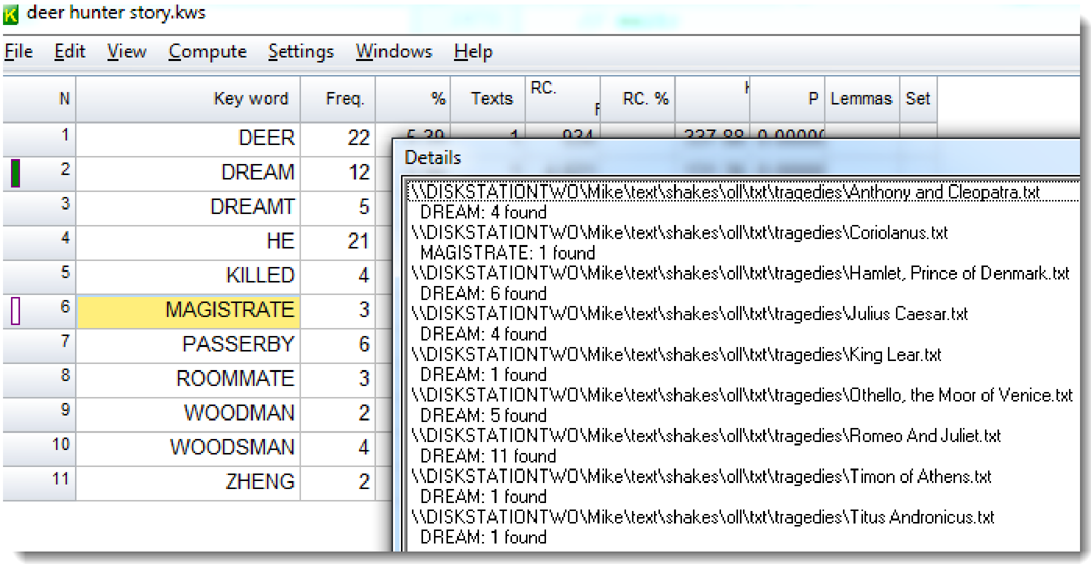
Suppose you have identified *muscle, fibre, protein* as key words in a specific text. You might want to find out whether there are any more texts in your corpus which use these words.

How to do it

This function can be reached in any window of data which contains the F menu option, e.g. a word-list or a key words 6 list.



It enables you to seek out all text files which contain **at least one** mention of the words you have marked or selected. Before you click, choose the set of texts 44 which you want to peruse. (If you haven't, the function will let you use the text(s) the current key words or word-list entries are based on.)



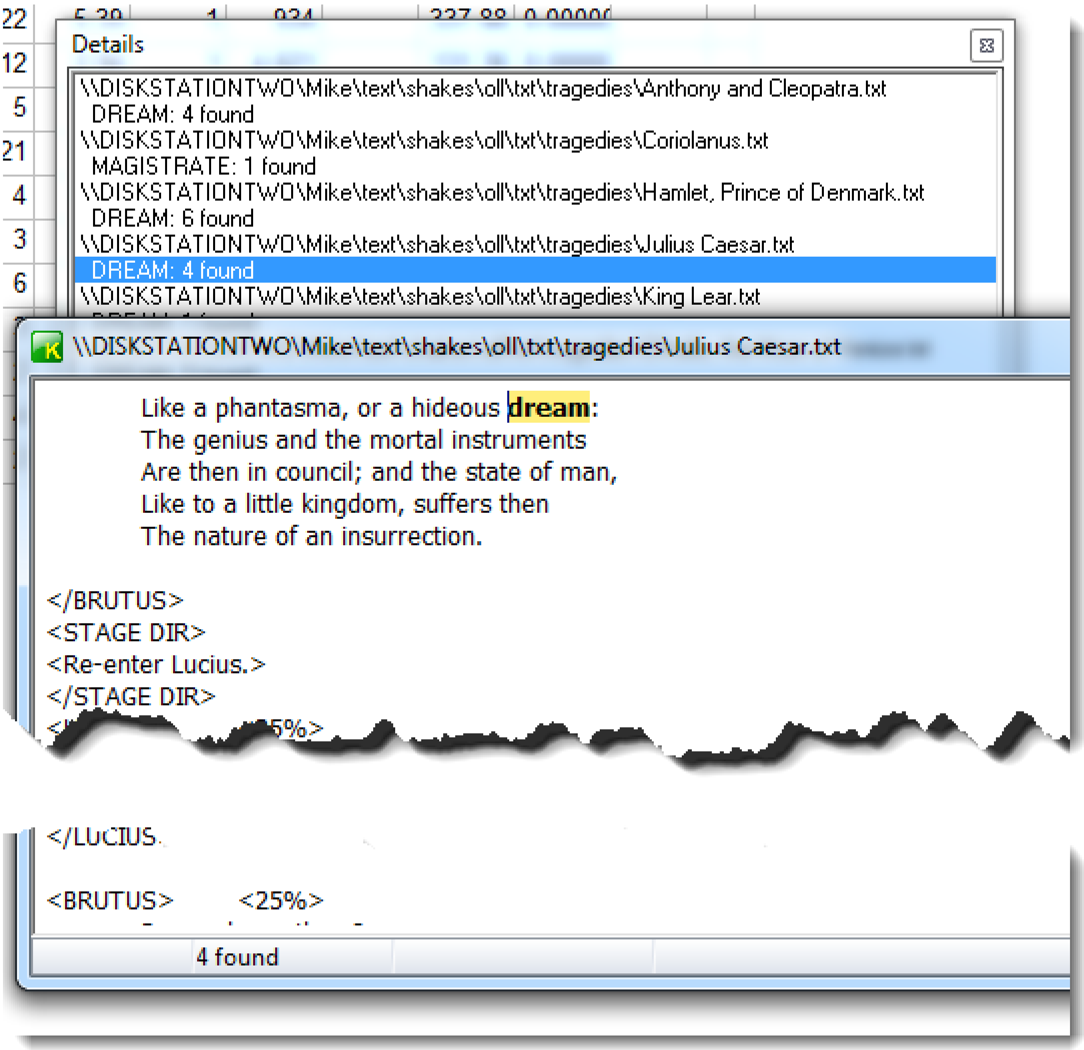
N	Key word	Freq.	%	Texts	RC. F	RC. %	K	P	Lemmas	Set
1	DEER	22	5.30	1	934		337.88	0.00000		
2	DREAM	12								
3	DREAMT	5								
4	HE	21								
5	KILLED	4								
6	MAGISTRATE	3								
7	PASSERBY	6								
8	ROOMMATE	3								
9	WOODMAN	2								
10	WOODSMAN	4								
11	ZHENG	2								

```
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Anthony and Cleopatra.txt
  DREAM: 4 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Coriolanus.txt
  MAGISTRATE: 1 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Hamlet, Prince of Denmark.txt
  DREAM: 6 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Julius Caesar.txt
  DREAM: 4 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\King Lear.txt
  DREAM: 1 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Othello, the Moor of Venice.txt
  DREAM: 5 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Romeo And Juliet.txt
  DREAM: 11 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Timon of Athens.txt
  DREAM: 1 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Titus Andronicus.txt
  DREAM: 1 found
```

Here we have a keywords list based on a Chinese folk tale with two items chosen by marking 112. The text files to examine in this case are all the Shakespeare tragedies...

What you get

A display based on all the words you marked, showing which text files 430 they were found in and how many of each word were found. If you double-click as shown



\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Anthony and Cleopatra.txt
DREAM: 4 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Coriolanus.txt
MAGISTRATE: 1 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Hamlet, Prince of Denmark.txt
DREAM: 6 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Julius Caesar.txt
DREAM: 4 found
\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\King Lear.txt

\\DISKSTATIONTWO\Mike\text\shakes\oll\txt\tragedies\Julius Caesar.txt

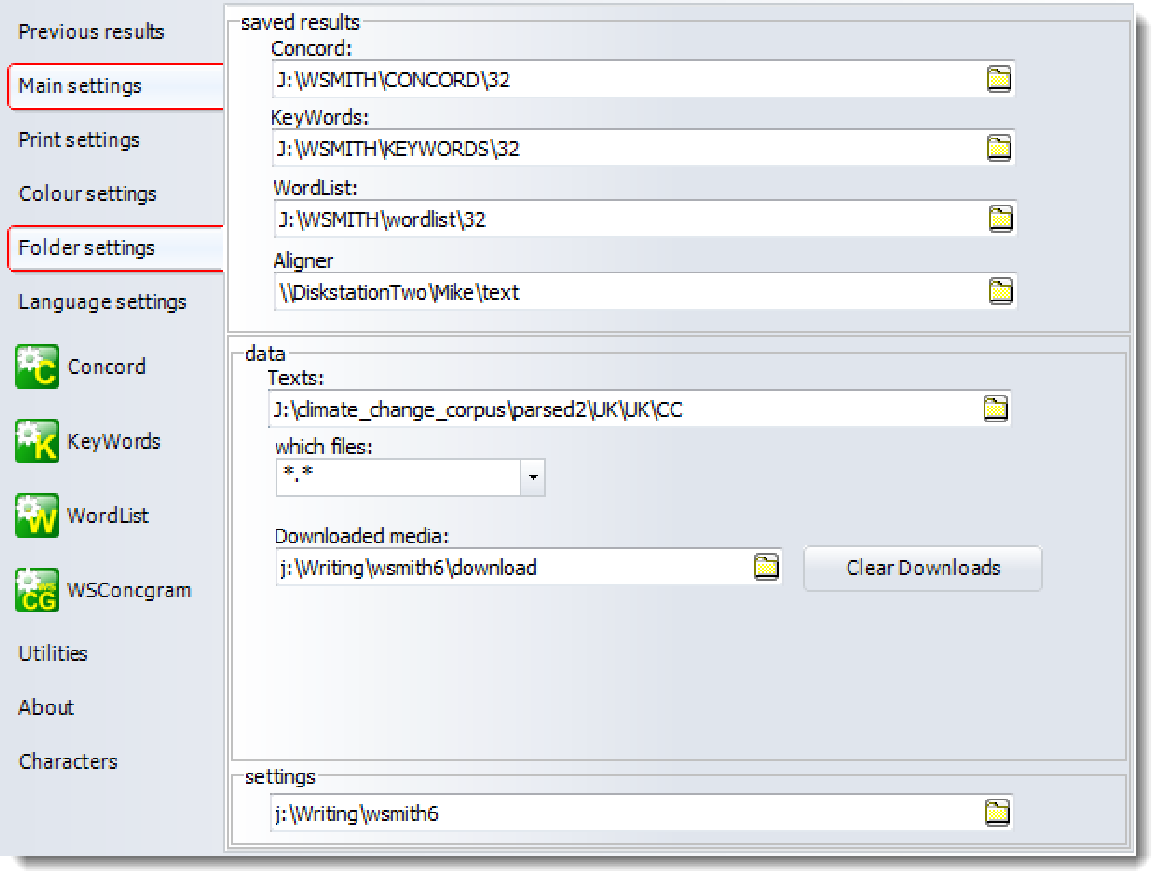
```
          Like a phantasma, or a hideous dream:
          The genius and the mortal instruments
          Are then in council; and the state of man,
          Like to a little kingdom, suffers then
          The nature of an insurrection.

</BRUTUS>
<STAGE DIR>
<Re-enter Lucius.>
</STAGE DIR>

</LUCIUS.

<BRUTUS>      <25%>
```

4 found

you'll get to see the source text and can examine each of the words, in this case the four tokens of the type *dream*.

5.18 folder settings

These are found in the main Controller.



saved results

Concord: J:\WSMITH\CONCORD\32

KeyWords: J:\WSMITH\KEYWORDS\32

WordList: J:\WSMITH\wordlist\32

Aligner \\DiskstationTwo\Mike\text

data

Texts: J:\climate_change_corpus\parsed2\UK\UK\CC

which files: *.*

Downloaded media: j:\Writing\wsmith6\download

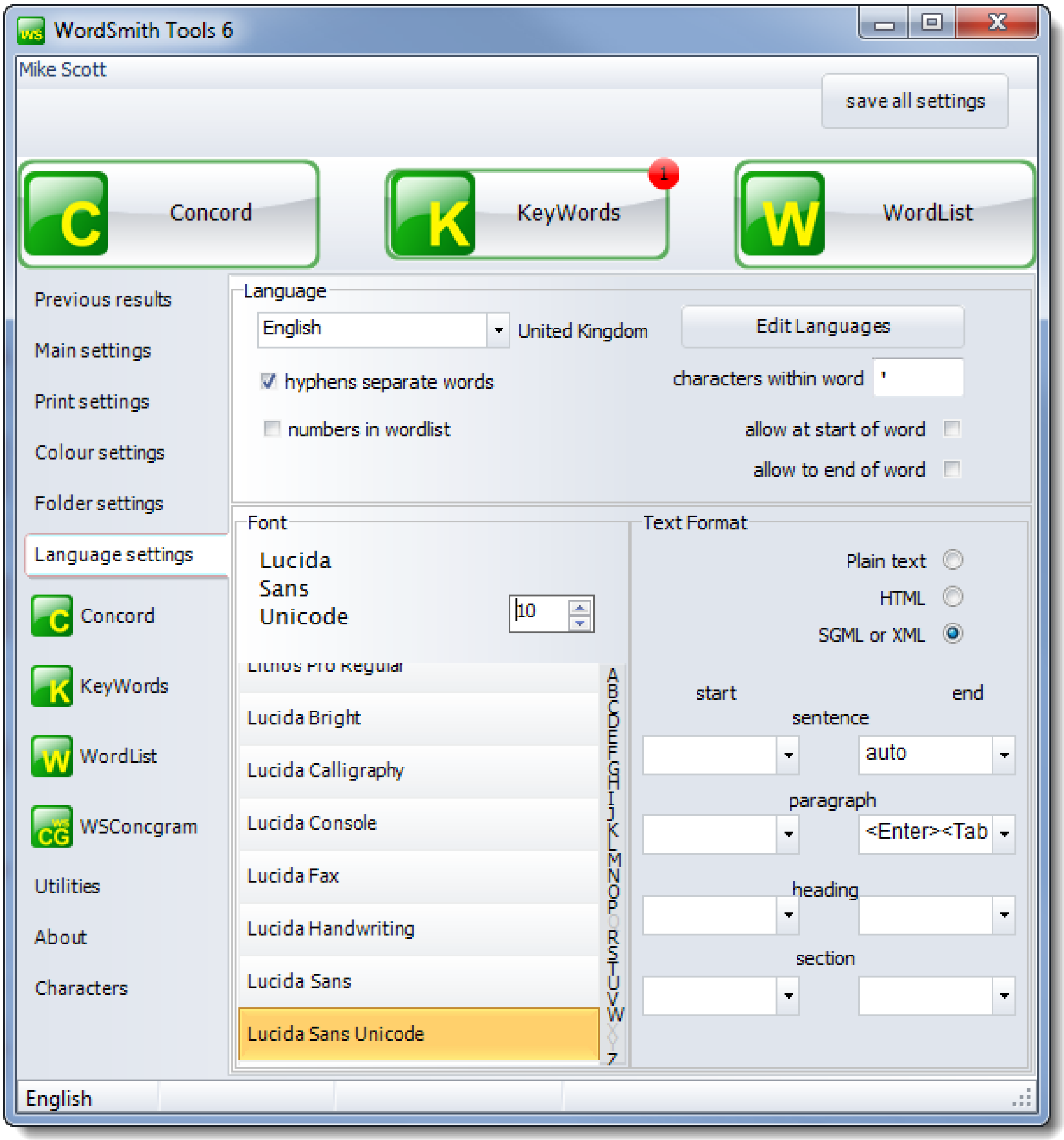
settings

j:\Writing\wsmith6

The settings folder will be default be a sub-folder of your My Documents folder but it can be set elsewhere if preferred.

5.19 fonts

Found by choosing *Settings* | *Font* in all Tools or via *Language Settings* in the Controller 4.
Enables you to choose a preferred Windows font and point size for the display windows and printing 97 in all the WordSmith Tools suite. Note that each language 81 can have its own different default font.



Language

English United Kingdom

☑ hyphens separate words

□ numbers in wordlist

characters within word '

allow at start of word □

allow to end of word □

Font

Lucida Sans Unicode 10

Text Format

Plain text ○

HTML ○

SGML or XML ●

	start	end
sentence		auto
paragraph		<Enter><Tab
heading		
section		

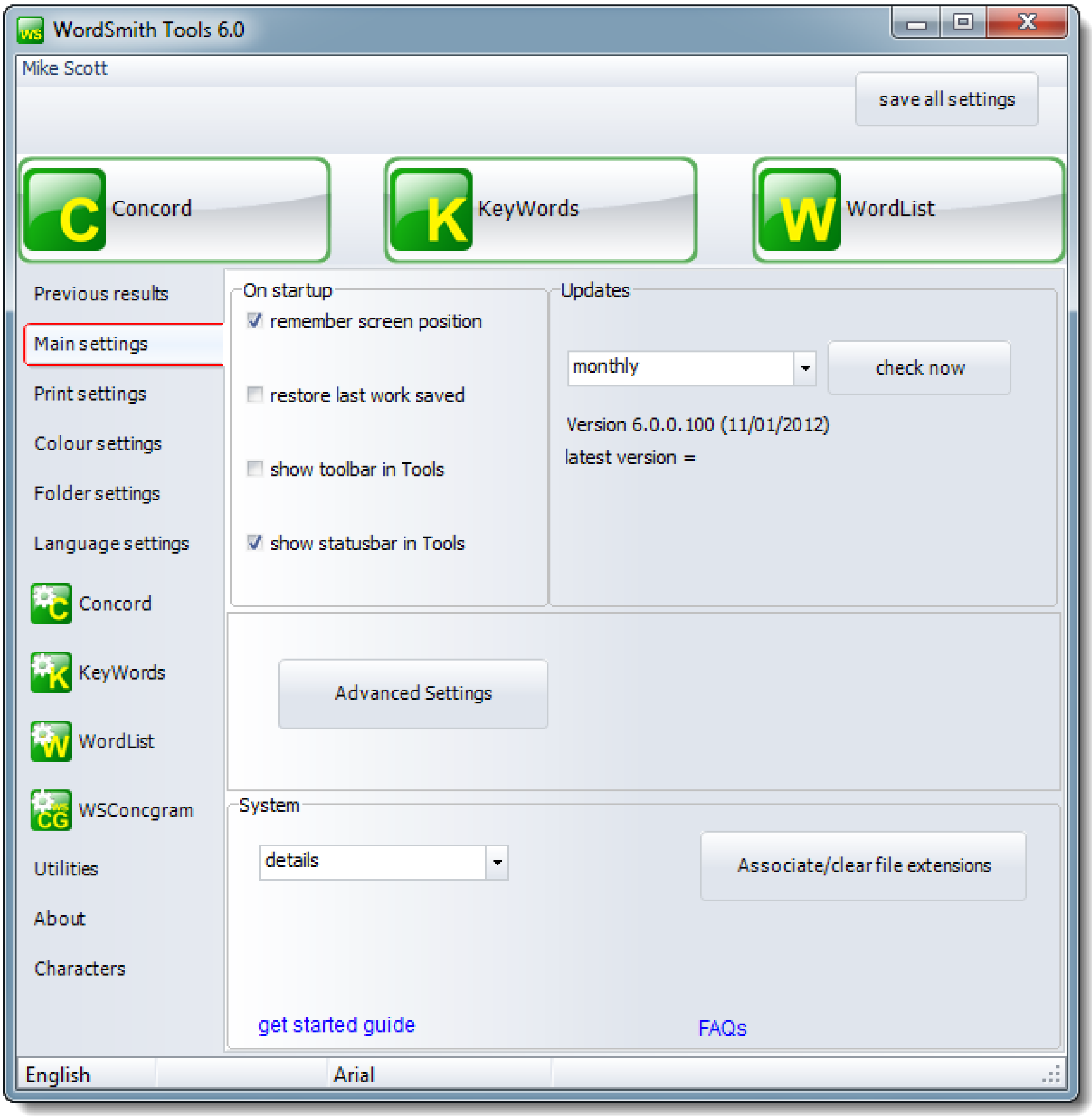
If you have data visible in any Tool, the font will automatically change; if you don't want any specific windows of data to change, because you want different font sizes or different character sets in different windows, minimise these first.

To set a column of data to bold, italics, underline etc., use the layout 87 option .

WordSmith Tools will offer fonts to suit the language 81 chosen in the top left box. Each language may require a special set of fonts. Language choice settings once saved can be seen (and altered, with care) in **`Documents\wsmith6\language_choices.ini`**.

5.20 main settings

Found in *Main settings* in the WordSmith Tools Controller 4.



WordSmith Tools 6.0

Mike Scott

On startup

☑ remember screen position

□ restore last work saved

□ show toolbar in Tools

☑ show statusbar in Tools

Updates

monthly

Version 6.0.0.100 (11/01/2012)

latest version =

System

details

English | Arial

Startup

Restore last work will bring back the last word-list, concordance or key-words list when you start WordSmith.

Show Help file will call up the Help file automatically when you start WordSmith.

Associate/clear file extensions will teach Windows to use (or not to use) Concord, WordList, KeyWords etc. to open the relevant files made by WordSmith.

Check for updates

WordSmith can be set to check for updated versions weekly, monthly or not at all. You may freely update your version within the version purchased (e.g. 6.0 allows you to update any 6.x version until 7.0 is issued).

Toolbar & Status bar

Each Tool has a status bar at the bottom and a toolbar with buttons at the top. By default the toolbar is hidden to reduce screen clutter.

System

The first box gives a chance to force the boxes which appear for choosing a file to show the files in various different ways. For example "details" will show listings with column headers so with one click you can order them by date and pick the most recent one even if you cannot remember the exact filename.

The *Associate/clear file extensions* button will teach Windows to use (or not to use) Concord, WordList, KeyWords etc. to open the relevant files made by WordSmith.

5.21 language

The point of it ...

1. Different languages sometimes require specific fonts.

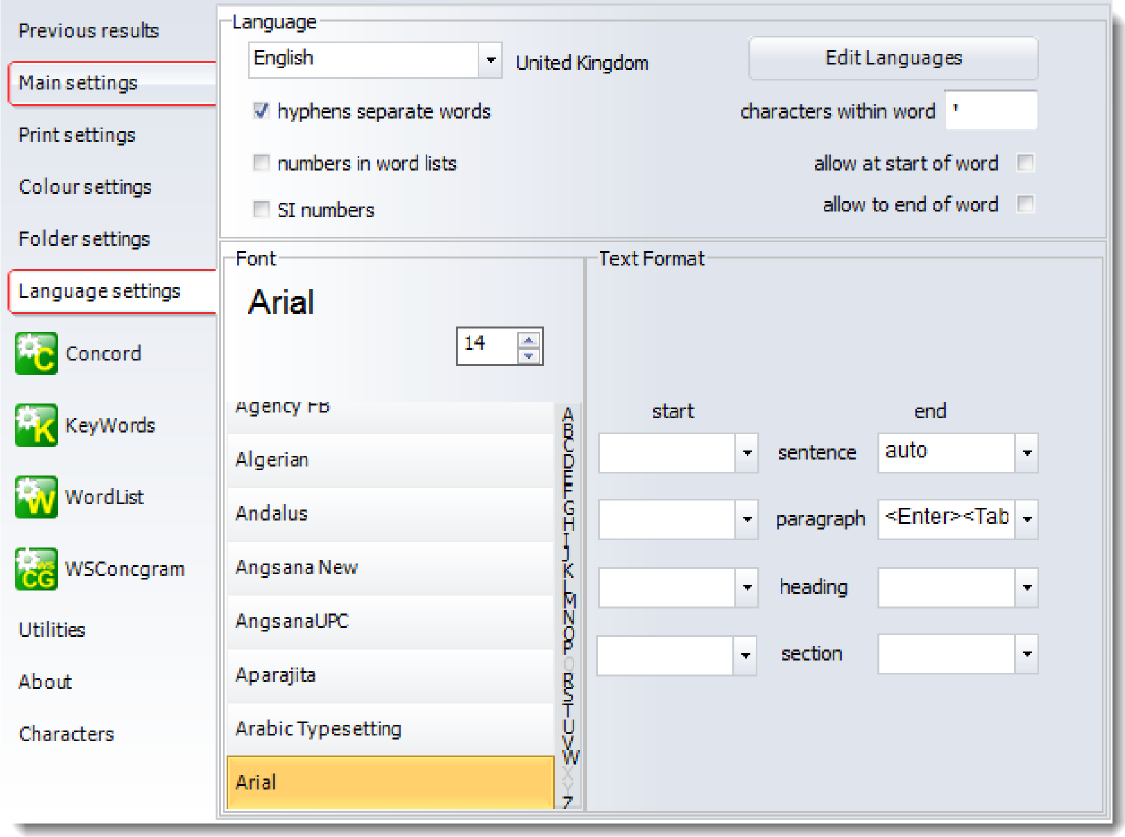
2. Languages vary considerably in their preferences regarding sorting order. Spanish, for example, uses this order: A,B,C,CH,D,E,F,G,H,I,J,K,L,LL,M,N,Ñ,O,P,Q,R,S,T,U,V,W,X,Y,Z. And accented characters are by default treated as equivalent to their unaccented counterparts in some languages (so, in French we get `donne, donné, données, donner, donnez`, etc.) but in other languages accented characters are not considered to be related to the unaccented form in this way (in Czech we get `cesta .. cas .. hre .. chodník` ..)

Sorting is handled using Microsoft routines. If you process texts in a language which Microsoft haven't got right, you should still see word-lists in a consistent order.

Note that case-sensitive means that `Mother` will come after `mother` (not before `apple` or after `zebra`).

It is important to understand that a comparison of two word-lists (e.g. in KeyWords) relies on sort order to get satisfactory results -- you will get strange results in this if you are comparing 2 word-lists which have been declared to be in different languages.

Settings



Language

English United Kingdom

☑ hyphens separate words

characters within word '

□ numbers in word lists

allow at start of word □

□ SI numbers

allow to end of word □

Font

Arial 14

Text Format

start		end
	sentence	auto
	paragraph	<Enter><Tab
	heading	
	section	

Choose the language for the text you're analysing in the Controller 4 under *Language Settings*. The language and character set 419 must be compatible, e.g. English is compatible with Windows Western (1252), DOS Multilingual (850).

WordSmith Tools handles a good range of languages, ranging from Albanian to Zulu. Chinese, Japanese, Arabic etc. are handled in Unicode. You can view word lists, concordances, etc. in different languages at the same time.

Hyphens separate words, Numbers, Characters within word 124

Font 78

Text Format 124

How more languages are added

Press the *Edit Languages* button.

See also: Choosing Accents & Symbols 420, Accented characters 419, Processing text in Chinese etc., Text Format 146, Changing language 419

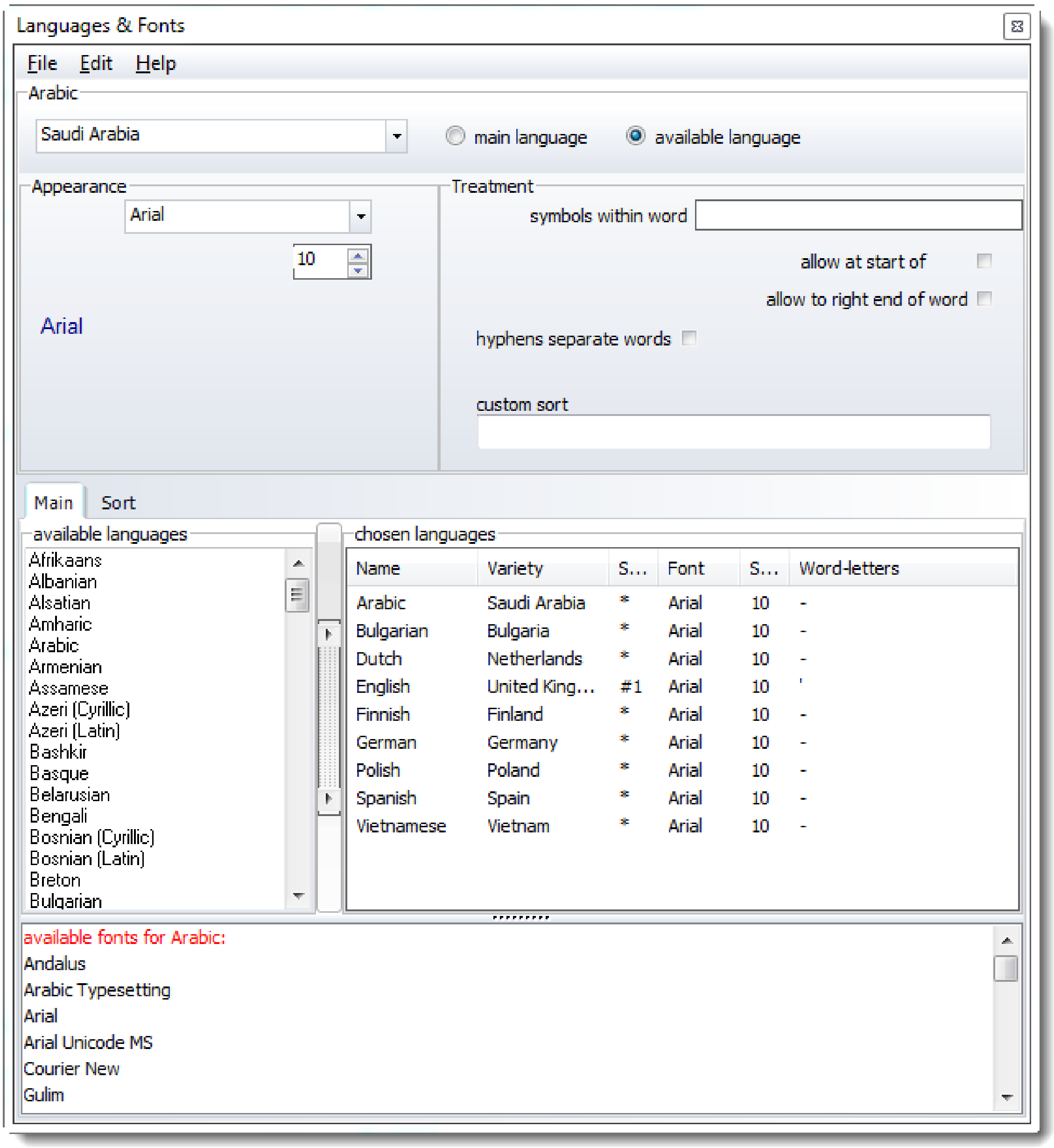
5.21.1 Overview

You will probably only need to do this once, when you first use WordSmith Tools.

How to get here

The Language Chooser is accessed from the main WordSmith Controller menu: *Settings* | *Main Settings* | *Text and Languages* | *Other Languages*.

What you will see may look like this:



Arabic

Saudi Arabia ○ main language ● available language

Appearance

Arial

10

Arial

Treatment

symbols within word

allow at start of □

allow to right end of word □

hyphens separate words □

custom sort

available languages

Afrikaans
Albanian
Alsatian
Amharic
Arabic
Armenian
Assamese
Azeri (Cyrillic)
Azeri (Latin)
Bashkir
Basque
Belarusian
Bengali
Bosnian (Cyrillic)
Bosnian (Latin)
Breton
Bulgarian

chosen languages

Name	Variety	S...	Font	S...	Word-letters
Arabic	Saudi Arabia	*	Arial	10	-
Bulgarian	Bulgaria	*	Arial	10	-
Dutch	Netherlands	*	Arial	10	-
English	United King...	#1	Arial	10	'
Finnish	Finland	*	Arial	10	-
German	Germany	*	Arial	10	-
Polish	Poland	*	Arial	10	-
Spanish	Spain	*	Arial	10	-
Vietnamese	Vietnam	*	Arial	10	-

available fonts for Arabic:
Andalus
Arabic Typesetting
Arial
Arial Unicode MS
Courier New
Gulim

9 languages have been chosen already.

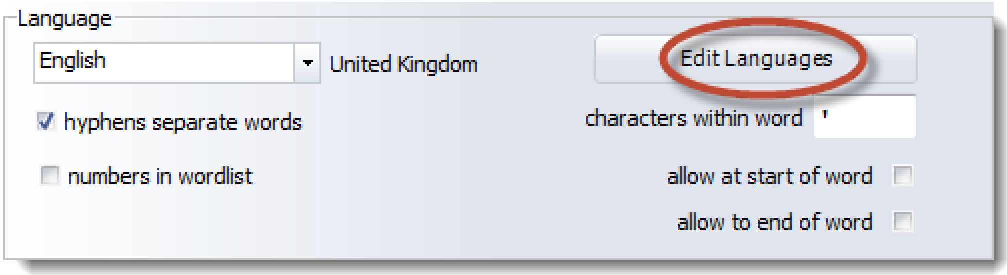
At the bottom you will see what the fonts on your system are for the current language selected.

See also : Language 84, Font 86, Sort Order 86, Other Languages 86, saving your choices 87, Changing the language of saved data 419

5.21.2 Language Chooser

How to get here

The Language Chooser is accessed from the main WordSmith Controller:



Language

English ▾ United Kingdom

☑ hyphens separate words

characters within word '

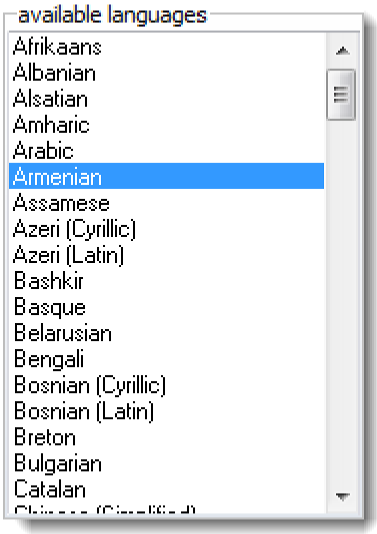
□ numbers in wordlist

allow at start of word □

allow to end of word □

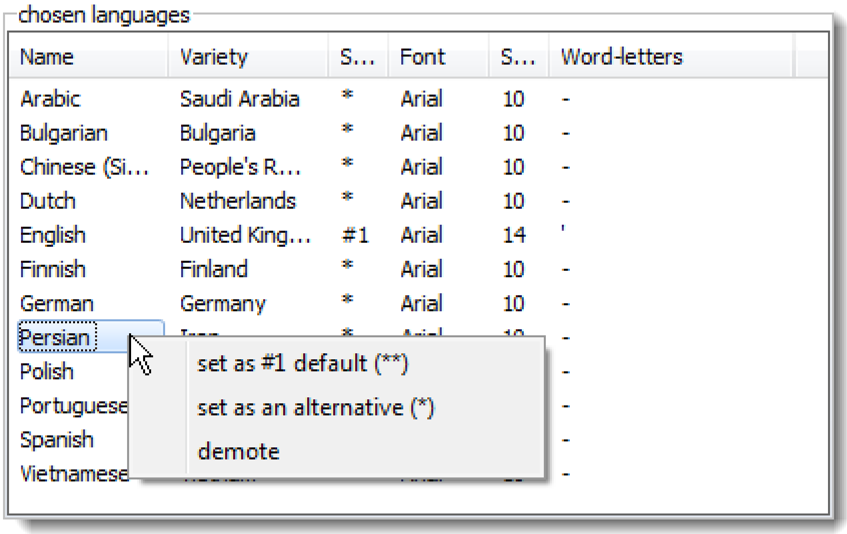
What it does

The list of languages on the left shows all those which are supported by the PC you're using. If any of them are greyed, that's because although they are "supported" by your version of Windows, they haven't been installed in your copy of Windows. (To install more multilingual support, you will need your original Windows cdrom or may be able to find help on the Internet.)



available languages

- Afrikaans
- Albanian
- Alsatian
- Amharic
- Arabic
- Armenian
- Assamese
- Azeri (Cyrillic)
- Azeri (Latin)
- Bashkir
- Basque
- Belarusian
- Bengali
- Bosnian (Cyrillic)
- Bosnian (Latin)
- Breton
- Bulgarian
- Catalan
- Chinese (Simplified)



chosen languages

Name	Variety	S…	Font	S…	Word-letters
Arabic	Saudi Arabia	*	Arial	10	-
Bulgarian	Bulgaria	*	Arial	10	-
Chinese (Si…	People's R…	*	Arial	10	-
Dutch	Netherlands	*	Arial	10	-
English	United King…	#1	Arial	14	'
Finnish	Finland	*	Arial	10	-
German	Germany	*	Arial	10	-
Persian	Iran	*	Arial	10	-
Polish					-
Portuguese					-
Spanish					-
Vietnamese					-

On the right, there are the currently chosen languages for use with WordSmith. The default language should be marked #1 and others which you might wish to use with *.

To change the status of a chosen language, right-click. This user is about to make Persian the #1 default. To delete any unwanted language, right-click and choose "demote". To add a language, drag it from the left window to the right, then set the country and font you prefer for that particular language.

For each chosen language, you can specify any symbols which can be included within a word, e.g. the apostrophe in English 125, where it makes more sense to think of "don't" as one word than as "don" and "t". You can also specify whether a hyphen separates words or not (e.g. whether "self-conscious" is to be considered as 2 words or 1).

Each time you change language, the list of fonts 86 available changes.

Some languages do not mark word-separators 427.

See also : Other Languages 86, saving your choices 87

5.21.3 Other Languages

To work on a language not in the list, go to New Language, type in a suitable name and base your new language on one of the existing languages. Choose a font which can show the characters & symbols you want to include. Sort order is handled as for the language you base your new language on.

See also : Language 84, Font 86, Sort Order 86, saving your choices 87

5.21.4 Font

The Fonts window shows those available for each language, depending on fonts you have installed. You will need a font which can show the characters you need: there are plenty of specialised fonts to be found on the Internet. Unicode fonts can show a huge number of different characters, but require your text to be saved in Unicode format. If you change font, the list of characters available changes.

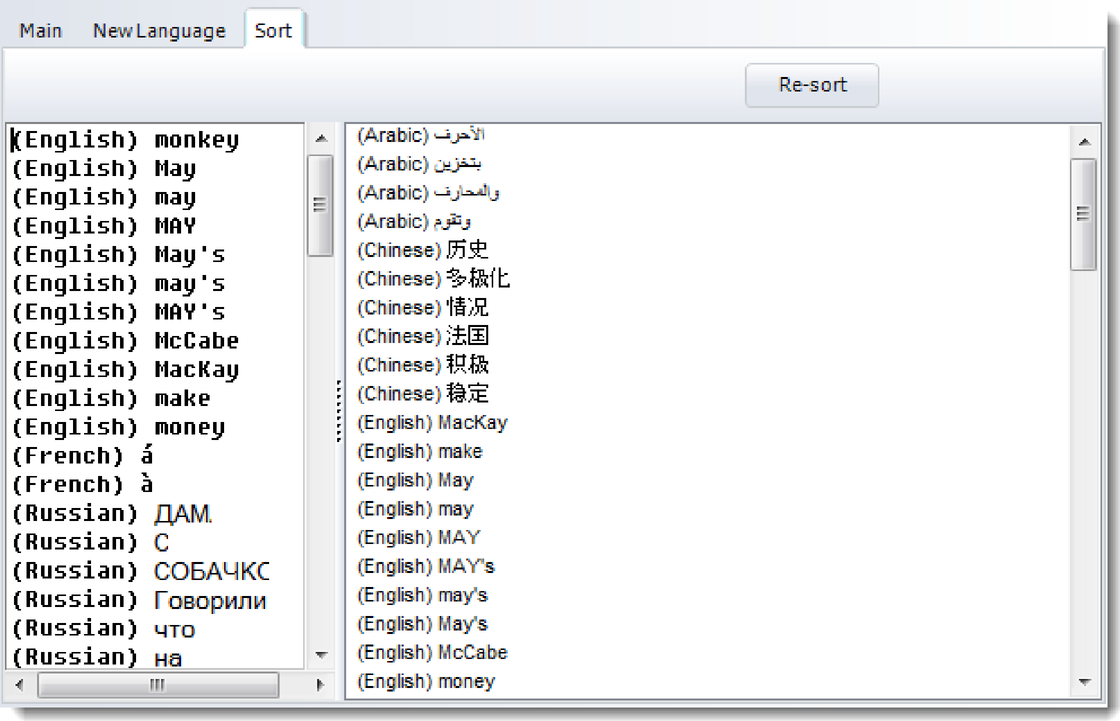
Click here for more on Unicode.

See also : Language 84, Sort Order 86, Other Languages 86, saving your choices 87

5.21.5 Sort Order

Sorting is done in accordance with the language chosen. (Spanish, Danish, etc. sort differently from English.)

The display



```
(English) monkey
(English) May
(English) may
(English) MAY
(English) May's
(English) may's
(English) MAY's
(English) McCabe
(English) MacKay
(English) make
(English) money
(French) á
(French) à
(Russian) ДАМ
(Russian) С
(Russian) СОБАЧКС
(Russian) Говорили
(Russian) что
(Russian) на
```

```
(Arabic) الأحرف
(Arabic) بتخزين
(Arabic) والمحارف
(Arabic) وتقوم
(Chinese) 历史
(Chinese) 多极化
(Chinese) 情况
(Chinese) 法国
(Chinese) 积极
(Chinese) 稳定
(English) MacKay
(English) make
(English) May
(English) may
(English) MAY
(English) MAY's
(English) may's
(English) May's
(English) McCabe
(English) money
```

- You will see 2 windows below "Resort" -- the one at the left contains some words in various languages; you can add your own. If your keyboard won't let you type them in, paste from your own collection of texts.
- The one at the right shows how these words get sorted according to the language you have selected.

See also : Language 84, Font 86, Other Languages 86, saving your choices 87

5.21.6 saving your choices

Save your results before quitting, so that next time WordSmith Tools will know your preferences regarding fonts and your #1 default language and your subsidiary default languages and you won't need to run this again. Results will be in **Documents\wsmith6\language_choices.ini**.

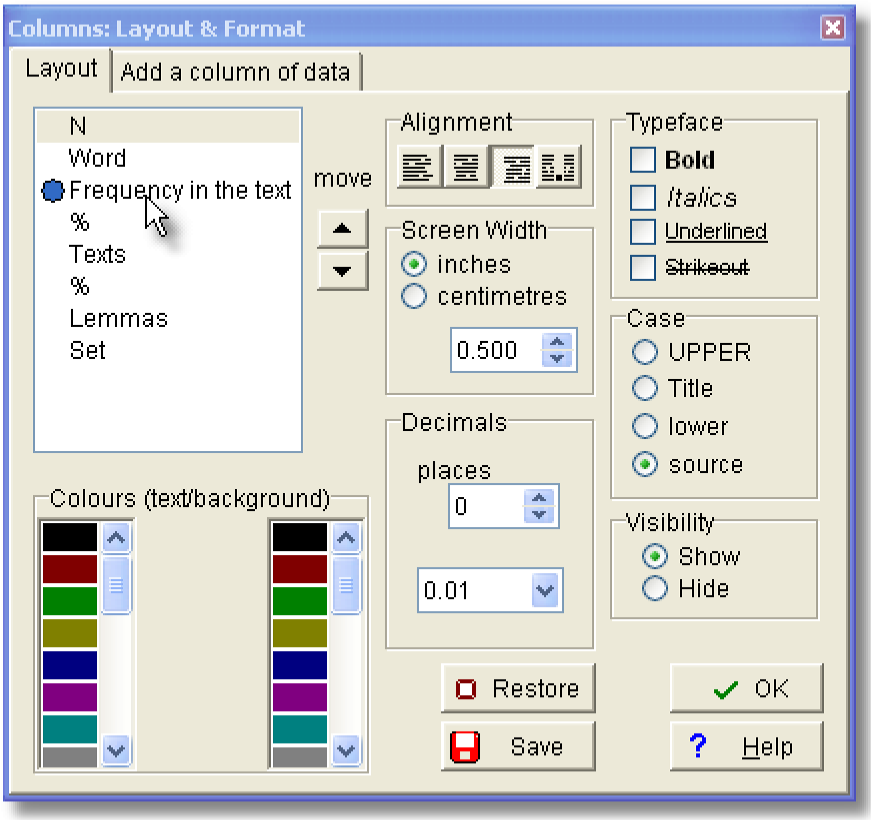
See also : Language 84, Font 86, Sort Order 86, Other Languages 86

5.22 layout & format

With any list open, right-click or choose View | Layout ▥ to choose your preferred display formats for each column of data.

Layout or Add data?

The *Layout* tab gives you a chance to format the layout of your data. *Add a column of data* lets you compute a new variable 63.



Alignment

Typeface: □ Bold □ Italics □ Underlined □ ~~Strikeout~~

Screen Width: ● inches ○ centimetres 0.500

Case: ○ UPPER ○ Title ○ lower ● source

Decimals: places 0; 0.01

Visibility: ● Show ○ Hide

You can edit the headings 71 by double-clicking and typing in your own preferred heading. "Frequency in the text" is too long but serves to illustrate.

Move

Click on the arrows to move a column up or down so as to display it in an alternative order.

Alignment

Allows a choice of left-aligned, centred, right-aligned, and decimal aligned text in each column, as appropriate.

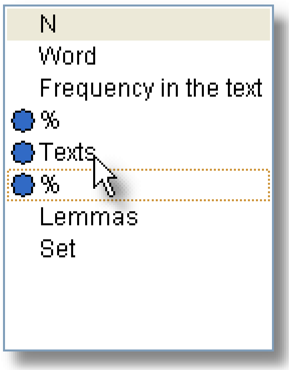
Typeface

Normal, bold, italic and/or underlined text. If none are checked, the typeface will be normal.

Screen Width

in your preferred units (cm. or inches).

Here 3 of the headings have been activated (by clicking) so that settings can be changed so as to get them all the same width.



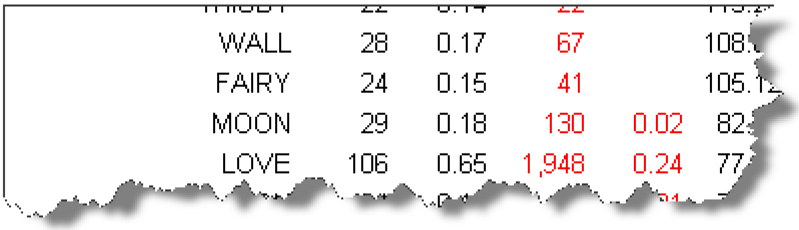
N
Word
Frequency in the text
● %
● Texts
● %
Lemmas
Set

Case

lower case, UPPER CASE, Title Case or source: as it came originally in the text file. The default for most data is upper case.

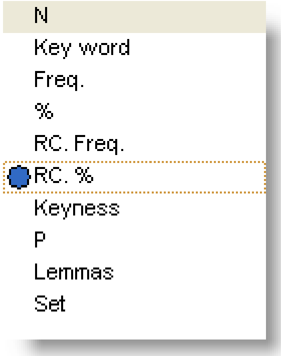
Decimals

the number of decimal places for numerical data, where applicable. For example, suppose you have this list of the key words of Midsummer Night's Dream in view but want to show the numbers in the column above 0.02, corresponding to WALL, FAIRY etc.,



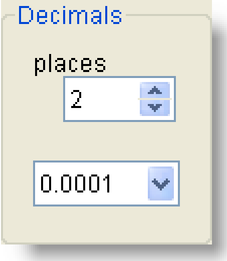
WALL	28	0.17	67		108.
FAIRY	24	0.15	41		105.1
MOON	29	0.18	130	0.02	82.
LOVE	106	0.65	1,948	0.24	77.

select the column(s) you want to affect,



N
Key word
Freq.
%
RC. Freq.
● RC. %
Keyness
P
Lemmas
Set

and set the decimals eg. like this

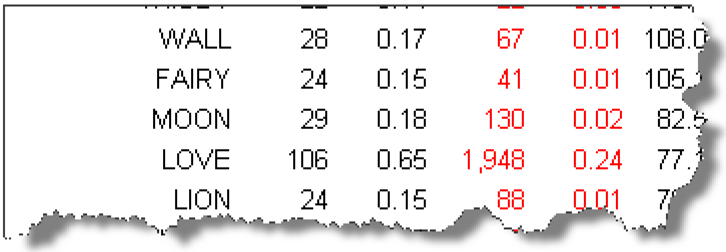


Decimals

places 2

0.0001

where the top number is the decimal places (2, unchanged from the default for percentage data) and the bottom is the threshold below which the data are not shown. In this case, any date smaller than 0.0001 won't be shown (the space will be blank). As soon as you make the change, you should immediately see the result.



WALL	28	0.17	67	0.01	108.0
FAIRY	24	0.15	41	0.01	105.
MOON	29	0.18	130	0.02	82.5
LOVE	106	0.65	1,948	0.24	77.
LION	24	0.15	88	0.01	7

Visibility

show or hide, or show only if greater than a certain number. (If this shows ***, then this option is not applicable to the data in the currently selected column.)

Colours

The bottom left window shows the available colours for the foreground & background. Click on a colour to change the display for the currently selected column of information.

Restore

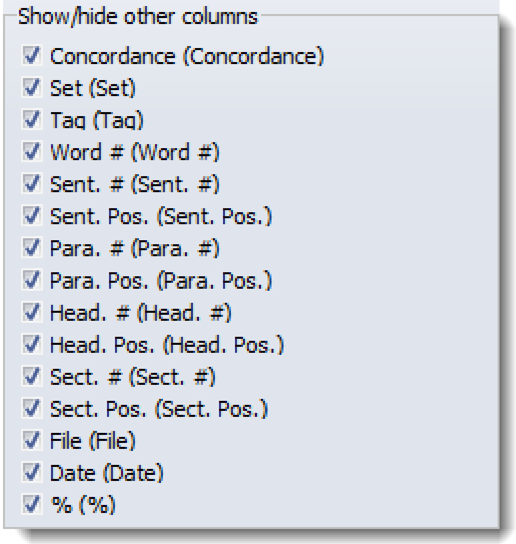
Restores settings to the state they were in before. Offers a chance to delete any custom saved layout for the current type of data (see Save).

Save

The point of this Save option is to set all future lists of the same type as the one you're working on to a preferred layout. Suppose you have a concordance open. If you change the layout as you like and save 101 the concordance in the usual way it will remember your settings anyway. But the next time you make a concordance, you'll get the WordSmith default layout. If you choose this Save, the next time you make a concordance, it will look like the current one.

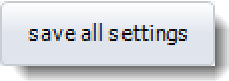
And a custom saved layout will be found in your **`Documents\wsmith6`** folder, eg. **`Concordance list customised.dat.`** **(**The only way of removing such settings would be to rename or delete that file.)

Alternatively you can choose always to show or hide certain columns of data with settings. For example, in the Controller's Concord settings the *What you see* tab offers these options,



Show/hide other columns

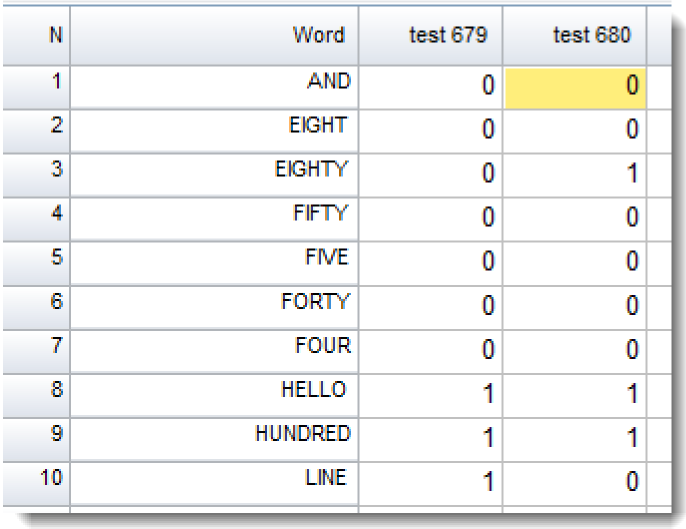
☑ Concordance (Concordance)
☑ Set (Set)
☑ Tag (Tag)
☑ Word # (Word #)
☑ Sent. # (Sent. #)
☑ Sent. Pos. (Sent. Pos.)
☑ Para. # (Para. #)
☑ Para. Pos. (Para. Pos.)
☑ Head. # (Head. #)
☑ Head. Pos. (Head. Pos.)
☑ Sect. # (Sect. #)
☑ Sect. Pos. (Sect. Pos.)
☑ File (File)
☑ Date (Date)
☑ % (%)



which can be saved permanently with .

Freeze the columns

If you have a lot of detailed consistency files and wish to freeze the word column so as to see the words for every column of numbers, choose View | Freeze columns... This allows you to set the number of fixed columns for example to 2, and the display will look as it does here:



N	Word	test 679	test 680
1	AND	0	0
2	EIGHT	0	0
3	EIGHTY	0	1
4	FIFTY	0	0
5	FIVE	0	0
6	FORTY	0	0
7	FOUR	0	0
8	HELLO	1	1
9	HUNDRED	1	1
10	LINE	1	0

where the N and Word columns are both frozen (and cannot be re-sorted) allowing you to look at the 679th and 680th text file data.

Similarly a statistics list allows the text file-names column to be frozen:

See also: setting & saving defaults 113, setting colour 60 choices in WordSmith Tools Controller 4 .

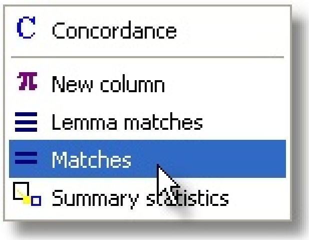
5.23 match words in list

The point of it…

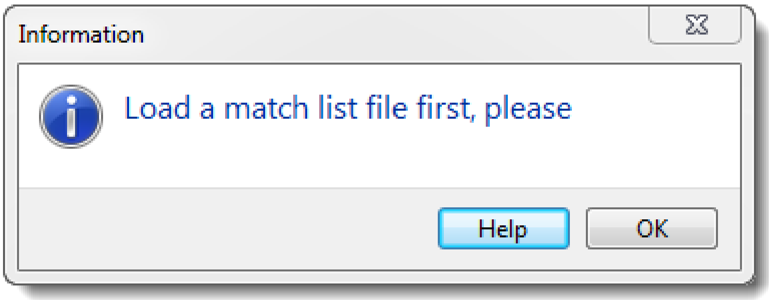
This function helps you filter your listing. You may choose to relate the entries in a concordance or list of words (word-list, collocate list, etc.) with a set of specific words which interest you. For example, to mark all those words in your list which are function words, or all those which end in **-ing**. Those which match are marked with a tilde (**~**). With the entries marked, you can then choose to delete all the marked entries (or all the unmarked ones), or sort them according to whether they're marked or not.

How to do it: WordList example

With a word-list loaded up using WordList, click in the column whose data you want to match up. This will usually be one showing words, not numbers. Then choose *Compute* | *Matches*.

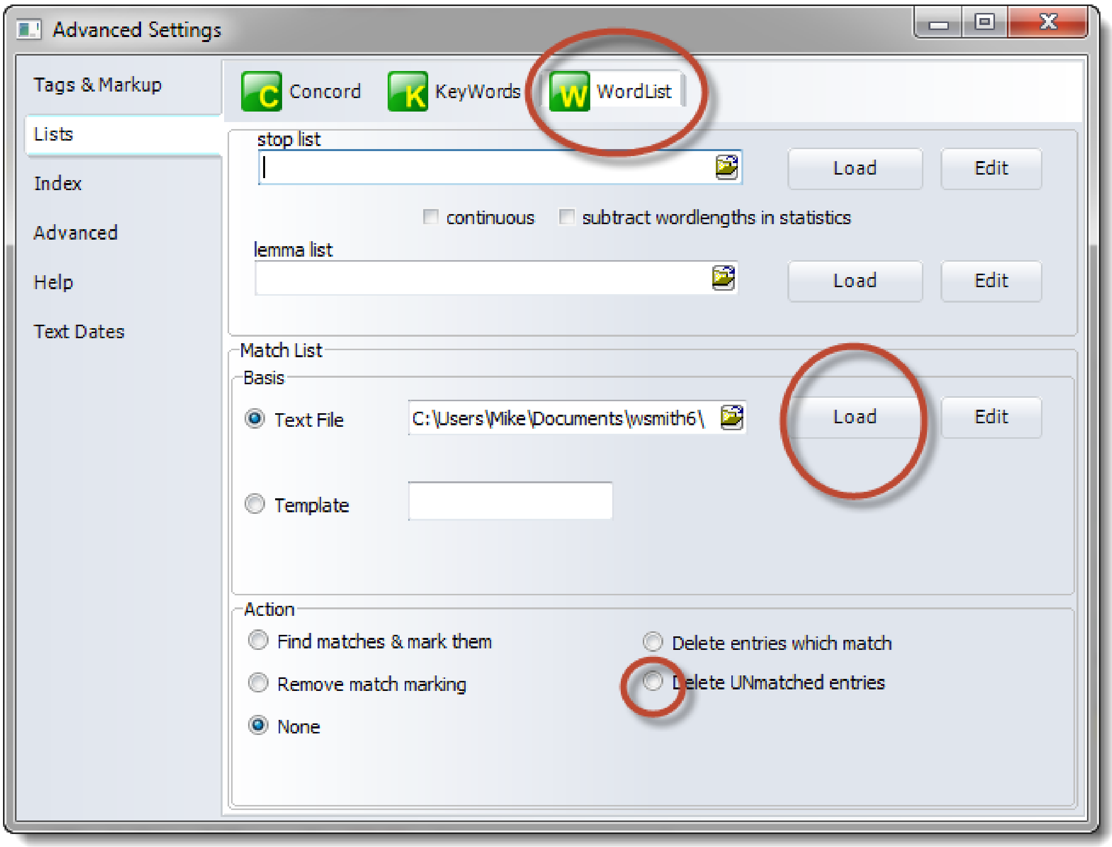


If you have no suitable match-list settings, you may get this:



Load a match list file first, please

The main Controller settings dialogue box appears.



Advanced Settings

stop list

□ continuous □ subtract wordlengths in statistics

lemma list

Match List

Basis

● Text File C:\Users\Mike\Documents\wsmith6\

○ Template

Action

○ Find matches & mark them

○ Remove match marking

● None

○ Delete entries which match

○ Delete UNmatched entries

The circled areas show some of the main choices: make sure you are choosing for the right Tool, and if matching words from a text file, browse to find it and then press *Load* to load its words. You must of course decide what is to be done with any matching entries.

Text File or Template

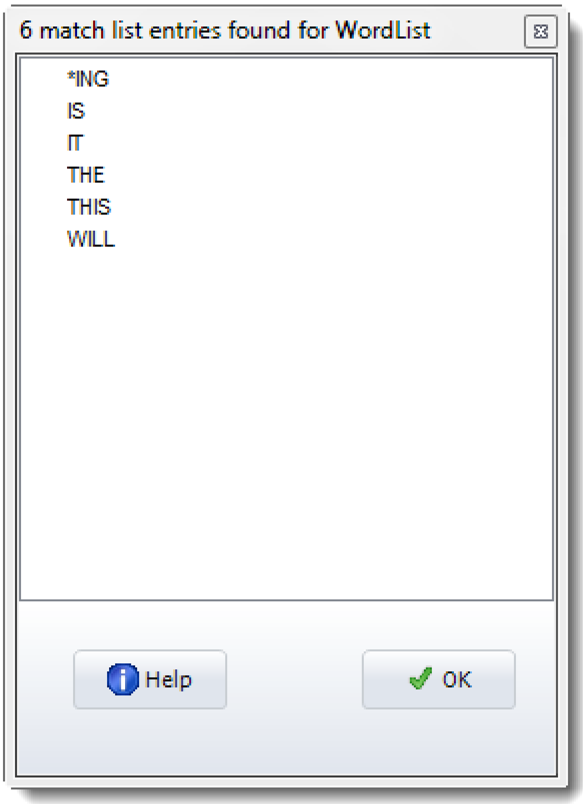
Choose now whether you want to filter by using a text file which contains all the words you're interested in (e.g. a plain text file of function words [not supplied]) or a template filter such as ***ing** (which checks every entry to see whether it contains a word ending in **ing**.).

To use a match list in a file, you first prepare a file, using **Notepad** or any plain text word processor, which specifies all the words you wish to match up. Separate each word using commas, or else place each one on a new line. You can use capital letters or lower-case as you prefer. You can use a semi-colon for comment lines. There is no limit to the number of words.

Example

```
; Match list for test purposes.
THE,THIS,IS
IT
WILL
*ING
```

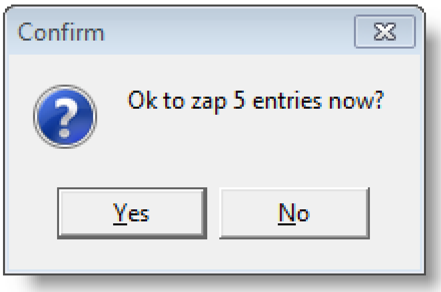
If you choose a file, the Controller will then read it and inform you as to how many words there are in it. (There is no limit to the number of words but only the first 50 will be shown in the Controller.)


6 match list entries found for WordList

*ING
IS
IT
THE
THIS
WILL

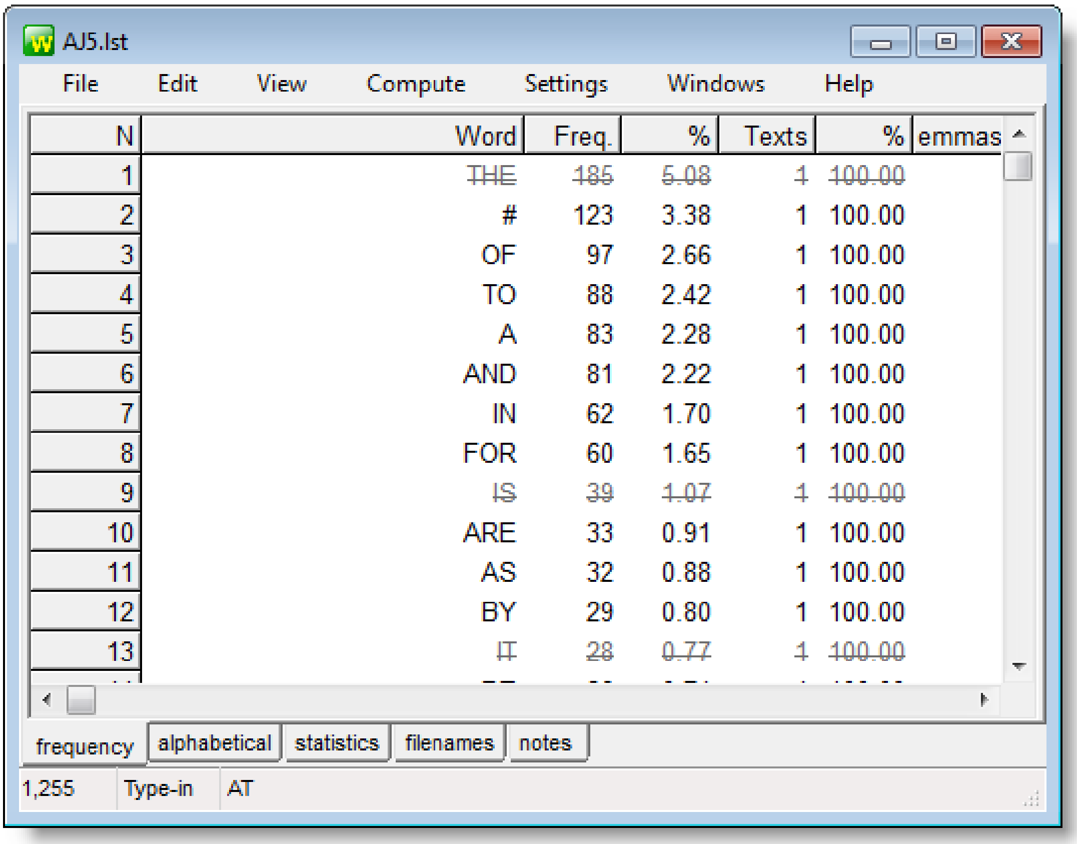
Action

The current Tool then checks every entry in the selected column in your current list to see whether it matches either the template or one of the words in your plain text file. Those which do match are marked or deleted as appropriate for the Action requested (as in the example below where five matching entries were found, the action selected was *delete entries which match* and the match list included **`THE`**, **`IS`** and **`IT`**).


Confirm

Ok to zap 5 entries now?

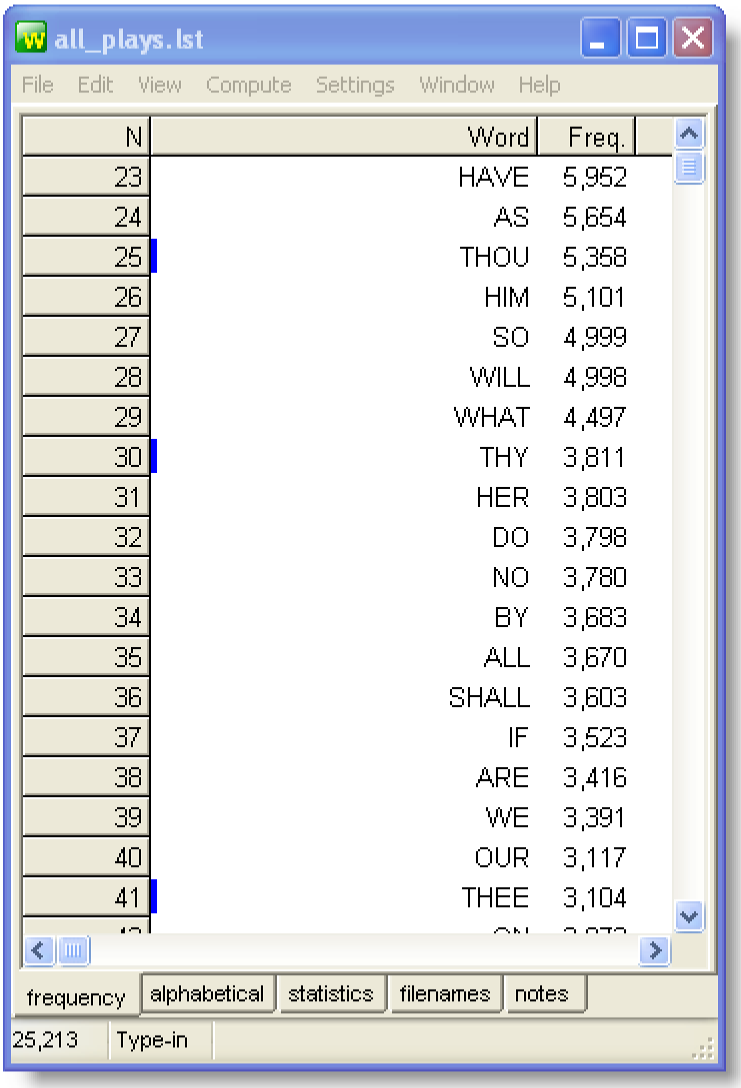
I answered No so you could see this result:



N	Word	Freq.	%	Texts	%	emmas
1	~~THE~~	~~185~~	~~5.08~~	~~1~~	~~100.00~~	
2	#	123	3.38	1	100.00	
3	OF	97	2.66	1	100.00	
4	TO	88	2.42	1	100.00	
5	A	83	2.28	1	100.00	
6	AND	81	2.22	1	100.00	
7	IN	62	1.70	1	100.00	
8	FOR	60	1.65	1	100.00	
9	~~IS~~	~~39~~	~~1.07~~	~~1~~	~~100.00~~	
10	ARE	33	0.91	1	100.00	
11	AS	32	0.88	1	100.00	
12	BY	29	0.80	1	100.00	
13	~~IT~~	~~28~~	~~0.77~~	~~1~~	~~100.00~~	

1,255 Type-in AT

In the screenshot below, the action was *find matches & mark them*, and the match-list contained archaic forms like **`thou, thee, thy`**.



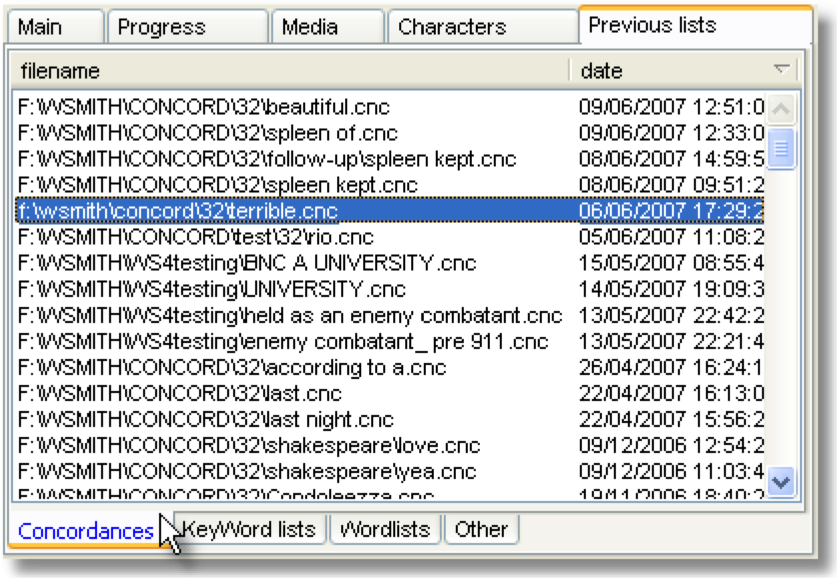
N	Word	Freq.
23	HAVE	5,952
24	AS	5,654
25	THOU	5,358
26	HIM	5,101
27	SO	4,999
28	WILL	4,998
29	WHAT	4,497
30	THY	3,811
31	HER	3,803
32	DO	3,798
33	NO	3,780
34	BY	3,683
35	ALL	3,670
36	SHALL	3,603
37	IF	3,523
38	ARE	3,416
39	WE	3,391
40	OUR	3,117
41	THEE	3,104

The marking can be removed using a menu option or by re-running the match-list function with *remove match marking* as the action.

You can obtain statistics of the matches, using the Summary Statistics 66 menu option.

See also: Comparing Word-lists 260, Comparing Versions 263, Stop Lists 120, Lemma Matching 274

5.24 previous lists



filename	date
F:\WSMITH\CONCORD\32\beautiful.cnc	09/06/2007 12:51:0
F:\WSMITH\CONCORD\32\spleen of.cnc	09/06/2007 12:33:0
F:\WSMITH\CONCORD\32\follow-up\spleen kept.cnc	08/06/2007 14:59:5
F:\WSMITH\CONCORD\32\spleen kept.cnc	08/06/2007 09:51:2
f:\wsmith\concord\32\terrible.cnc	06/06/2007 17:29:2
F:\WSMITH\CONCORD\test\32\rio.cnc	05/06/2007 11:08:2
F:\WSMITH\WS4testing\BNC A UNIVERSITY.cnc	15/05/2007 08:55:4
F:\WSMITH\WS4testing\UNIVERSITY.cnc	14/05/2007 19:09:3
F:\WSMITH\WS4testing\held as an enemy combatant.cnc	13/05/2007 22:42:2
F:\WSMITH\WS4testing\enemy combatant_ pre 911.cnc	13/05/2007 22:21:4
F:\WSMITH\CONCORD\32\according to a.cnc	26/04/2007 16:24:1
F:\WSMITH\CONCORD\32\last.cnc	22/04/2007 16:13:0
F:\WSMITH\CONCORD\32\last night.cnc	22/04/2007 15:56:2
F:\WSMITH\CONCORD\32\shakespeare\love.cnc	09/12/2006 12:54:2
F:\WSMITH\CONCORD\32\shakespeare\yea.cnc	09/12/2006 11:03:4
F:\WSMITH\CONCORD\32\Condoleezza.cnc	19/11/2006 18:40:2

These windows show the lists of results you have obtained in previous uses of WordSmith.

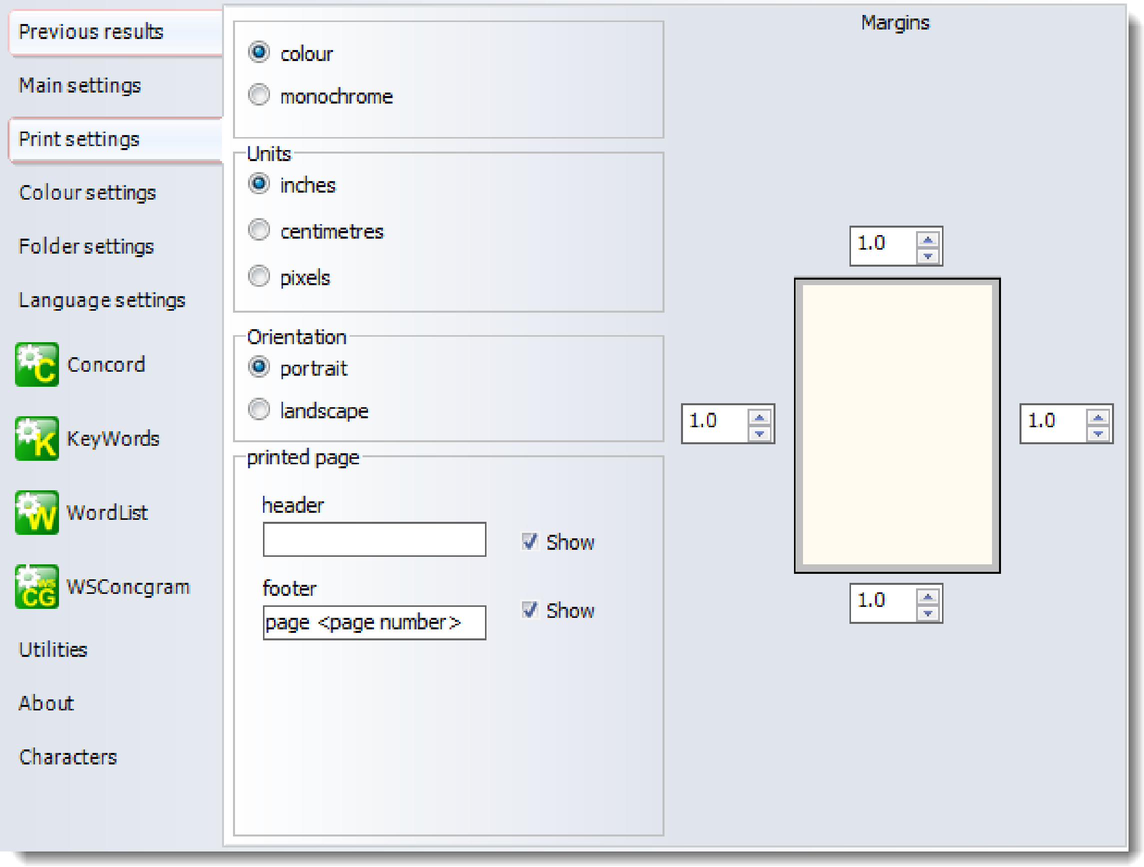
To see any of these, simply select it and double-click -- the appropriate Tool will be called up and the data shown in it.

The popup menu for the window is accessed by a right-click on your mouse.

To delete an entry, select it and then press *Del*.
To re-sort your entries click the header or choose *Resort* in the popup menu.

5.25 print and print preview

Print settings are in the main Controller:



● colour
○ monochrome

Units
● inches
○ centimetres
○ pixels

Orientation
● portrait
○ landscape

printed page

header

☑ Show

footer
page <page number>

☑ Show

Margins

1.0

1.0

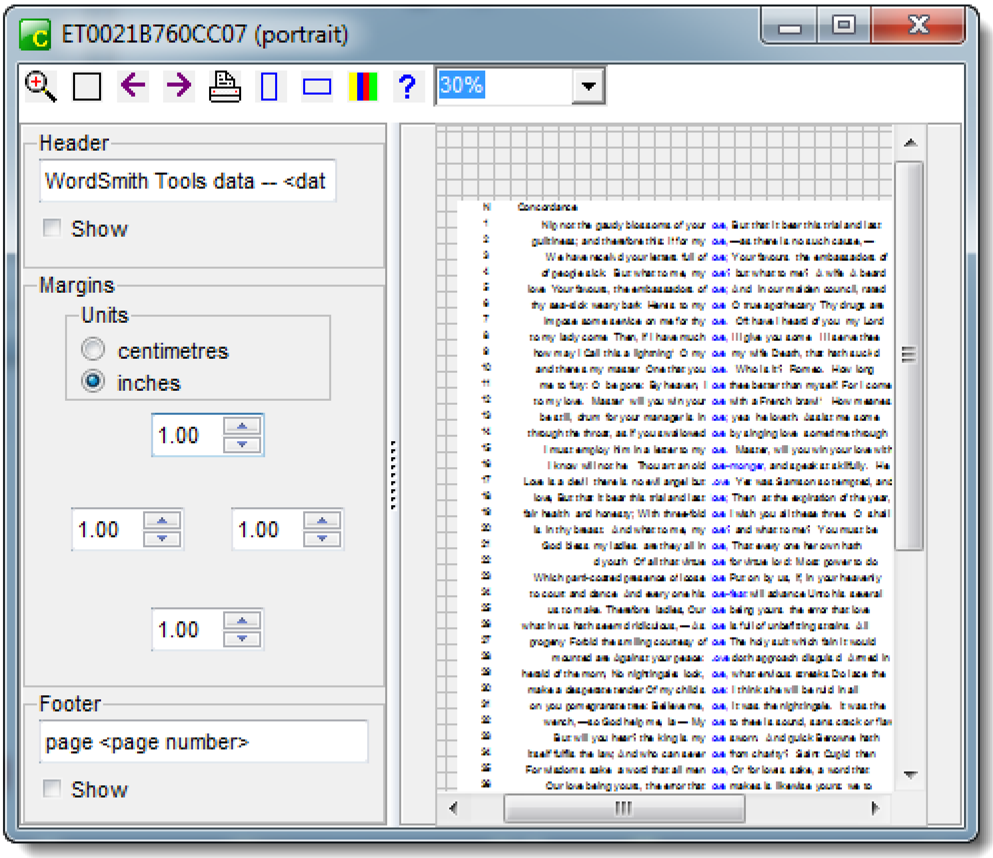
1.0

1.0

Print Settings

If you set printing to monochrome, your printer will use italics or bold type for any columns using other than the current "plain text" colour 60. Otherwise it will print in colour on a colour printer, or in shades of grey if the printer can do grey shading. You can also change the units, adjust orientation (portrait ▯ or landscape ▭) and margins and default header and footer.

When you choose a print or print preview menu item in a Tool, you'll be taken by default to a print preview, which shows you what the current page of data looks like, and from which you can print.



ET0021B760CC07 (portrait)

30%

Header

WordSmith Tools data -- <dat

□ Show

Margins

Units

○ centimetres

● inches

1.00

1.00 1.00

1.00

Footer

page <page number>

□ Show

Bigger and Smaller

Zoom to 100% (🔍) or fit to page (🔍), or choose a view in the list. The display here works in exactly the same way as the printing to paper. Any slight differences between what you see and what you get are due to font differences.

You can also pull the whole print preview window larger or smaller.

Next (→) & Last (←) Page

Takes you forward or back a page.

Portrait (▯) or Landscape (▭)?

Sets printing to the page shape you want.

Header, Footer, Margins

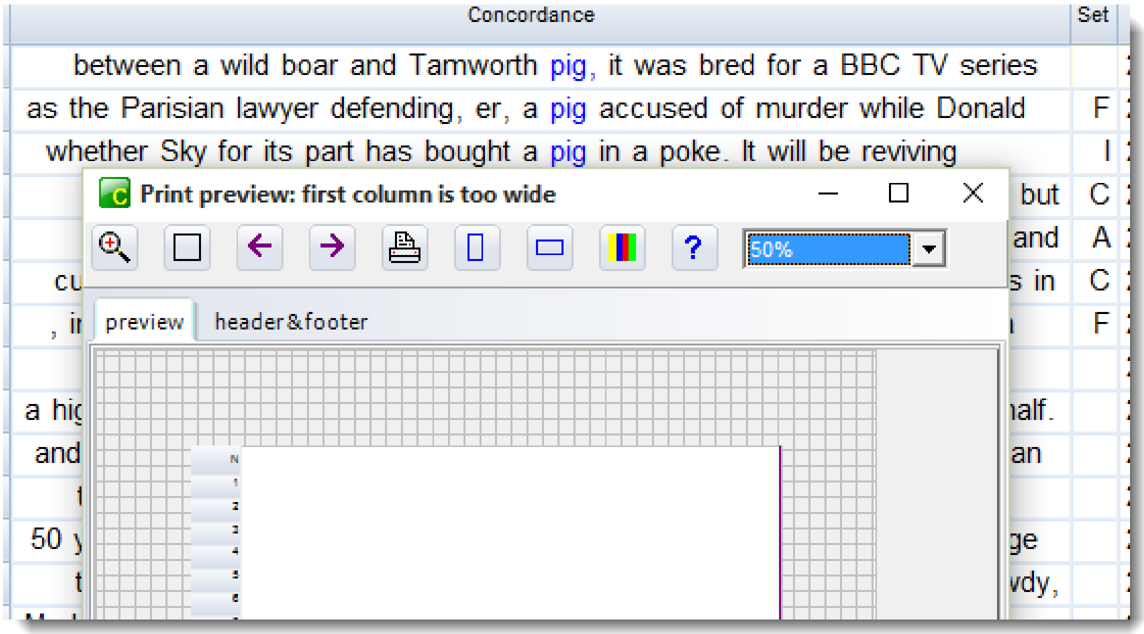
You can type a header & footer to appear on each page. Press *Show* if you want them included. If you include <date> this will put today's date and <page number> does the numbering. Margins are altered by clicking the numbers -- you will see the effect in the print previews space at the right.

Print (🖶)

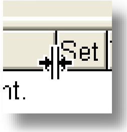
This calls up the standard Windows printer page and by default sets it to print the current page. You can choose other pages in this standard dialogue box if you want.

Some columns of data not shown

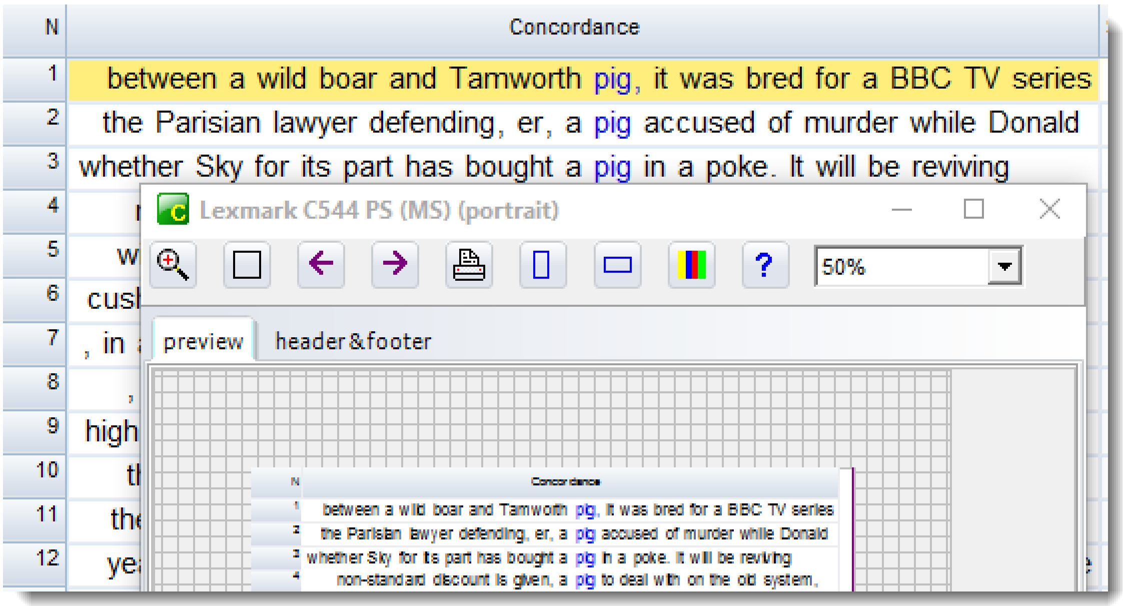
A case like this showing nothing but the line numbers



Concordance	Set
between a wild boar and Tamworth pig, it was bred for a BBC TV series	
as the Parisian lawyer defending, er, a pig accused of murder while Donald	F
whether Sky for its part has bought a pig in a poke. It will be reviving	I



is because you have pulled the concordance data too wide for the paper. WordSmith prints only any columns of data which are going to fit. Shrink the column, hide[87] any unwanted ones, or else set the print surface to landscape.



N	Concordance
1	between a wild boar and Tamworth pig, it was bred for a BBC TV series
2	the Parisian lawyer defending, er, a pig accused of murder while Donald
3	whether Sky for its part has bought a pig in a poke. It will be reviving
4	I
5	w
6	cus
7	, in
8	,
9	high
10	th
11	the
12	ye

N	Concordance
1	between a wild boar and Tamworth pig, it was bred for a BBC TV series
2	the Parisian lawyer defending, er, a pig accused of murder while Donald
3	whether Sky for its part has bought a pig in a poke. It will be reviving
4	non-standard discount is given, a pig to deal with on the old system,

See also: Printer Settings 80

5.26 quit WordSmith

Alt+X is the hot key.

Closing WordSmith Tools Controller 4 will close down all of the Tools.

If you press Alt+X, or use the System menu Close commands, you will get a chance to save any unsaved sets of data before the Tool in question closes. You will be asked to confirm closure if any window of data is still open.

If you're in a hurry, use the "no-check Exit" menu option which by-passes these checks.

By default, the last word list, concordance or key words listing that you saved or retrieved will be automatically restored on entry to WordSmith Tools. This feature can be turned off temporarily via a menu option or permanently in **`Documents\wsmith6\wordsmith6.ini`**.

5.27 saving

5.27.1 save results

To save your corrected results use *Save* (Ctrl+F2) in the menu. This saves all the results so you can return to the data at a later date. You may wish to clean up any deleted items by zapping 129, first.

Saved data is in a special **WordSmith Tools** format. The only point of it is to make it possible to use the data again another day. You will not be able to examine it usefully outside the Tools. If you want to export your data to a spreadsheet, graphics program, database or word processor, etc., you can do this either by saving as text 102 or by copying the data to the clipboard 422.

save part of the data only

By default, and save all your data that you haven't zapped 129 . If you want to save only part of it, but don't want to zap it to oblivion, choose Copy 66.

avoid the save prompts?

You can avoid them in the Advanced settings 36 (*Main Settings* | *Advanced Settings* | *Advanced* | *Other*).

5.27.2 save as text

The point of it...

Save as Text means save your data as a plain text file (as opposed to the WordSmith format for retrieving the data another day). It is usually quicker to copy selected text into the clipboard 422, e.g. if you simply want to insert your results into your word processor.

If you want to copy the data in colour, you should definitely use the clipboard 422.

In the case of a concordance, if you want only the words visible in your concordance line (not the number of characters mentioned below), use the clipboard and then Paste or Paste Special in graphics format.

How to do it

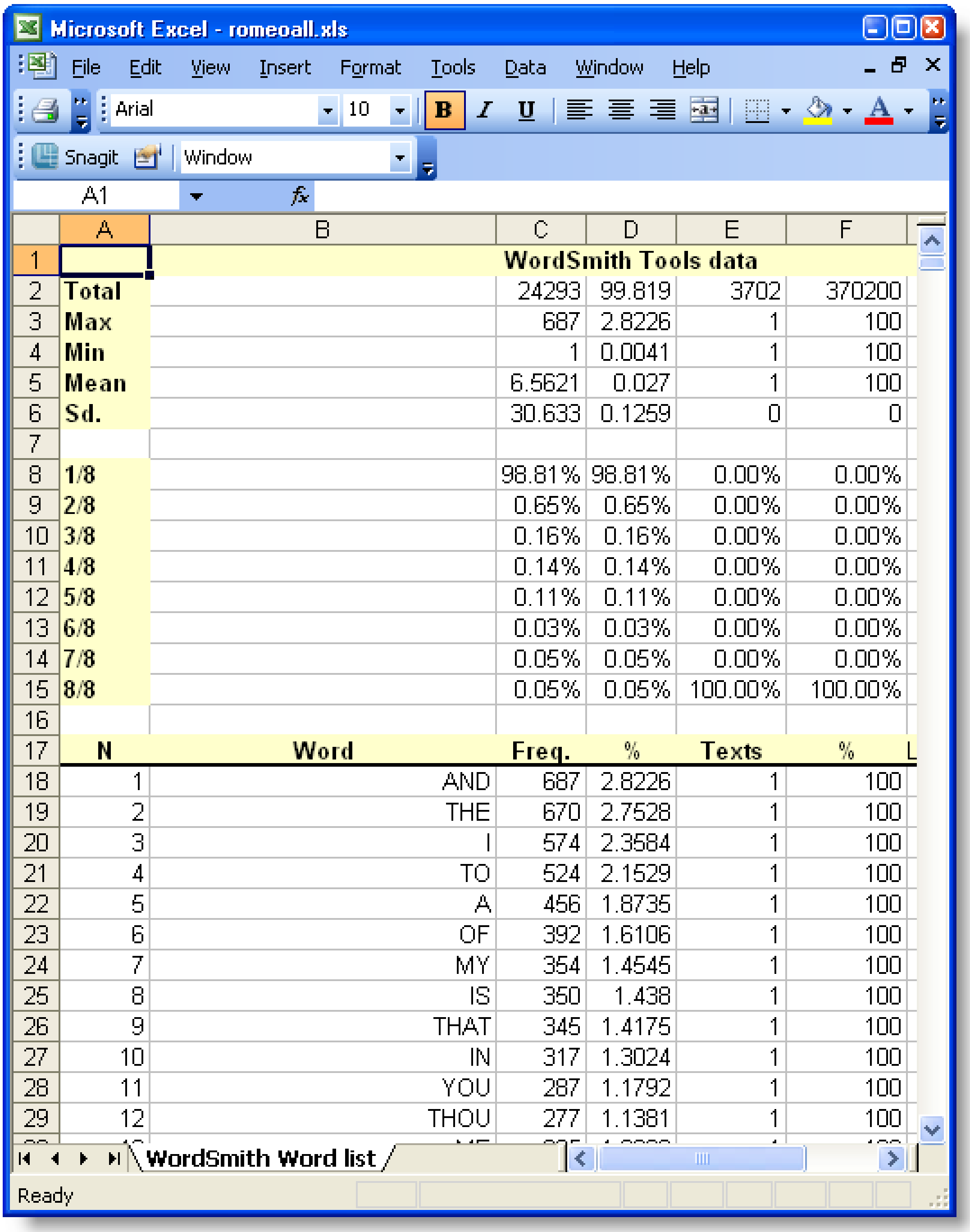
This function can be reached by Save As .. | Plain text (txt), XML text (), Excel spreadsheet () or Print to File (via F3 or) or Copy () to text file.

Options include:

header	words you want to save at the start of the data (leave blank if not wanted);
numbered	whether the numbers visible in the column at the left are saved too
column separator	by default a tab but you can specify something else to go between visible columns
rows	all/any which you have highlighted/a specific range, e.g. 1-10, 5-, -3
columns	all/any which you have highlighted/a specific range (column 1 is the one with the numbers)

You can then easily retrieve the data in your spreadsheet, database, word-processor, etc. (If you want to use it as a table in a word processor, first save as text, then in your word-processor choose the Convert Text to Table option if available. Choose to separate text at tabs.)

Note: The Excel spreadsheet () save will look something like this:



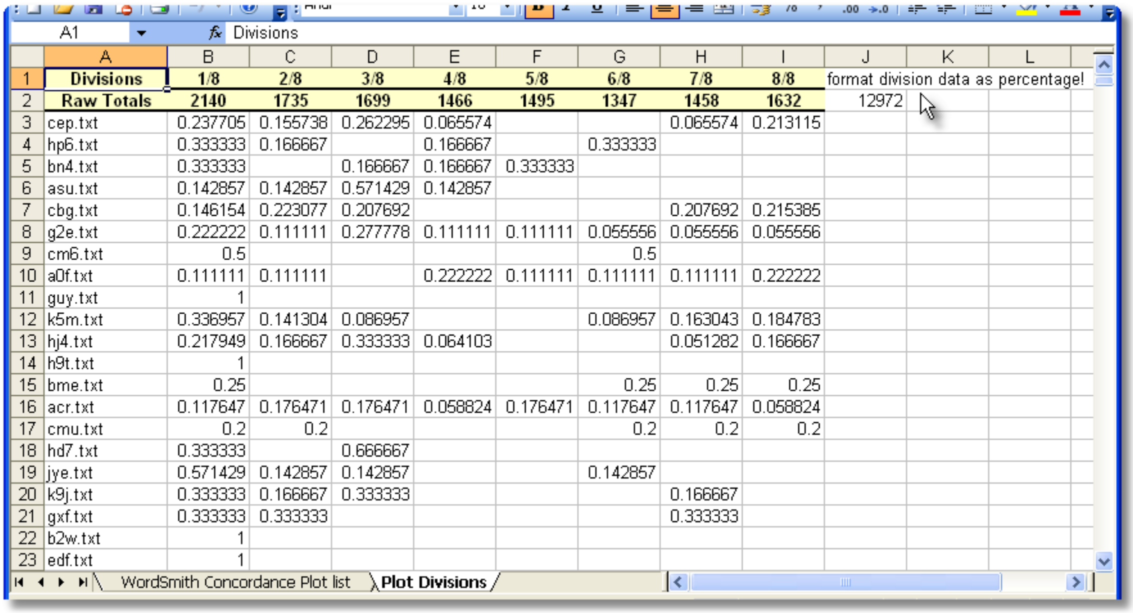
	A	B	C	D	E	F
1			WordSmith Tools data			
2	Total		24293	99.819	3702	370200
3	Max		687	2.8226	1	100
4	Min		1	0.0041	1	100
5	Mean		6.5621	0.027	1	100
6	Sd.		30.633	0.1259	0	0
7						
8	1/8		98.81%	98.81%	0.00%	0.00%
9	2/8		0.65%	0.65%	0.00%	0.00%
10	3/8		0.16%	0.16%	0.00%	0.00%
11	4/8		0.14%	0.14%	0.00%	0.00%
12	5/8		0.11%	0.11%	0.00%	0.00%
13	6/8		0.03%	0.03%	0.00%	0.00%
14	7/8		0.05%	0.05%	0.00%	0.00%
15	8/8		0.05%	0.05%	100.00%	100.00%
16						
17	N	Word	Freq.	%	Texts	%
18	1	AND	687	2.8226	1	100
19	2	THE	670	2.7528	1	100
20	3	I	574	2.3584	1	100
21	4	TO	524	2.1529	1	100
22	5	A	456	1.8735	1	100
23	6	OF	392	1.6106	1	100
24	7	MY	354	1.4545	1	100
25	8	IS	350	1.438	1	100
26	9	THAT	345	1.4175	1	100
27	10	IN	317	1.3024	1	100
28	11	YOU	287	1.1792	1	100
29	12	THOU	277	1.1381	1	100

The words are visible from row 18 onwards; above them we get some summary data. The 1/8, 2/8 etc. section splits the data into eighths; thus 100% of the Texts data (column E) is in the 8th section, whereas nearly all the data (98.8%) are in the smallest section in terms of word frequency, because so many words come once only. You'll be asked whether to compute this summary data if you choose to save as Excel.

In the case of a concordance line, saving as text will save as many "characters in 'save as text'" as you have set (adjustable in the Controller Concord Settings 222). The reason for this is that you will probably want a fixed number of characters, so that when using a non proportional font the search-words line up nicely. See also: Concord save and print 211.

Each worksheet can only handle up to 65,000 rows and 256 columns. If necessary there will be continuation sheets.

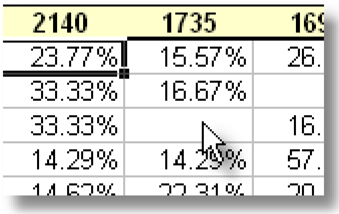
If your data contains a plot you will also get another worksheet in the Excel file, looking like this.



A1 Divisions

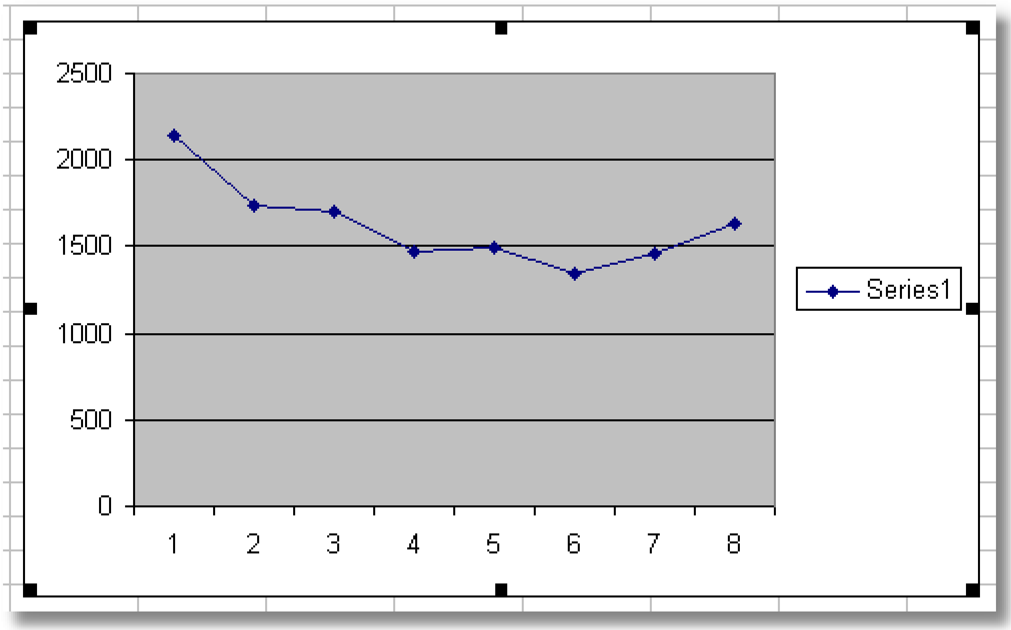
	A	B	C	D	E	F	G	H	I	J	K	L
1	**Divisions**	**1/8**	**2/8**	**3/8**	**4/8**	**5/8**	**6/8**	**7/8**	**8/8**	format division data as percentage!		
2	**Raw Totals**	**2140**	**1735**	**1699**	**1466**	**1495**	**1347**	**1458**	**1632**	12972		
3	cep.txt	0.237705	0.155738	0.262295	0.065574			0.065574	0.213115			
4	hp6.txt	0.333333	0.166667		0.166667		0.333333					
5	bn4.txt	0.333333		0.166667	0.166667	0.333333						
6	asu.txt	0.142857	0.142857	0.571429	0.142857							
7	cbg.txt	0.146154	0.223077	0.207692				0.207692	0.215385			
8	g2e.txt	0.222222	0.111111	0.277778	0.111111	0.111111	0.055556	0.055556	0.055556			
9	cm6.txt	0.5					0.5					
10	a0f.txt	0.111111	0.111111		0.222222	0.111111	0.111111	0.111111	0.222222			
11	guy.txt	1										
12	k5m.txt	0.336957	0.141304	0.086957			0.086957	0.163043	0.184783			
13	hj4.txt	0.217949	0.166667	0.333333	0.064103			0.051282	0.166667			
14	h9t.txt	1										
15	bme.txt	0.25					0.25	0.25	0.25			
16	acr.txt	0.117647	0.176471	0.176471	0.058824	0.176471	0.117647	0.117647	0.058824			
17	cmu.txt	0.2	0.2				0.2	0.2	0.2			
18	hd7.txt	0.333333		0.666667								
19	jye.txt	0.571429	0.142857	0.142857			0.142857					
20	k9j.txt	0.333333	0.166667	0.333333				0.166667				
21	gxf.txt	0.333333	0.333333					0.333333				
22	b2w.txt	1										
23	edf.txt	1										

The plot data are divided into the number of segments set for the ruler 441 (here they are eighths), and the percentage of each get put into the appropriate columns. That is, cell B3 means that 23.7% of the **`cep.txt`** data come in the first eighth of the text file. Set the format correctly as percentages in Excel, and you will see something like this:

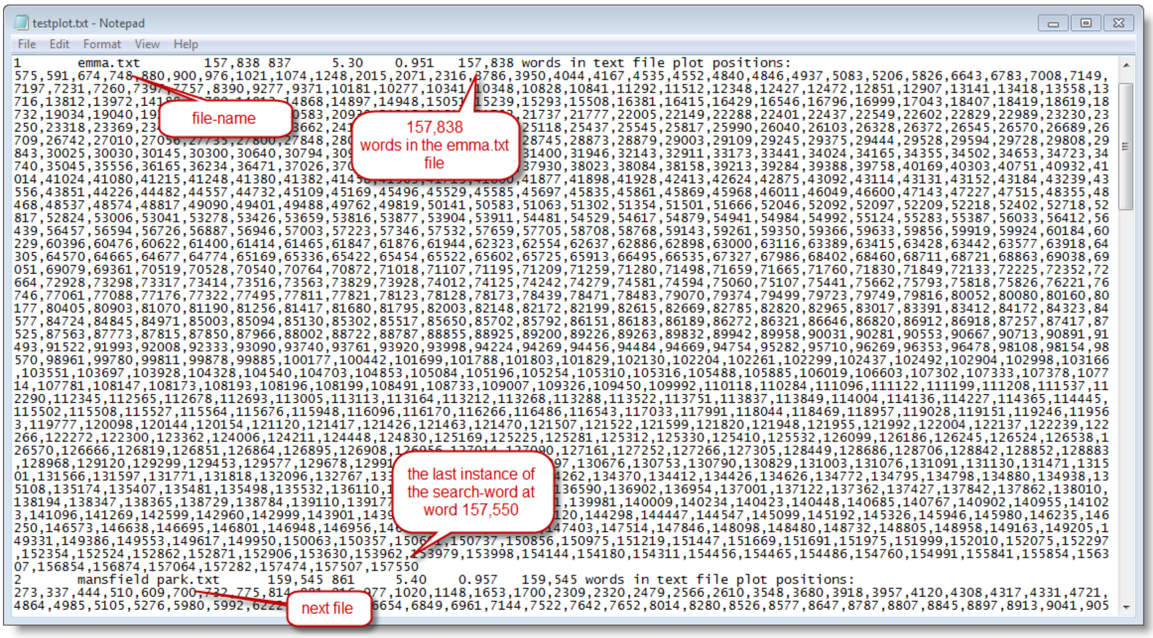


2140	**1735**	**169**
23.77%	15.57%	26.
33.33%	16.67%	
33.33%		16.
14.29%	14.29%	57.
14.62%	22.31%	20.

At the top you get the raw data, which you can use Excel to create a graphic with.



If you want access to the details of the plot, choose save as text. The results will look like this:



```
1       emma.txt        157,838 837     5.30    0.951   157,838 words in text file plot positions:
575,591,674,748,880,900,976,1021,1074,1248,2015,2071,2316,3786,3950,4044,4167,4535,4552,4840,4846,4937,5083,5206,5826,6643,6783,7008,7149,
7197,7231,7260,7397,7757,8390,9277,9371,10181,10277,10341,10348,10828,10841,11292,11512,12348,12427,12472,12851,12907,13141,13418,13558,13
716,13812,13972,141[illegible]14868,14897,14948,15051,15239,15293,15508,16381,16415,16429,16546,16796,16999,17043,18407,18419,18619,18
732,19034,19040,19[illegible]0583,2093[illegible]21737,21777,22005,22149,22288,22401,22437,22549,22602,22829,22989,23230,23
250,23318,23369,23[illegible]3662,241[illegible]25118,25437,25545,25817,25990,26040,26103,26328,26372,26545,26570,26689,26
709,26742,27010,27056,27735,27800,27848,280[illegible]28745,28873,28879,29003,29109,29245,29375,29444,29528,29594,29728,29808,29
843,30025,30030,30145,30300,30640,30794,309[illegible]31400,31946,32143,32911,33173,33441,34024,34165,34355,34502,34653,34723,34
740,35045,35556,36165,36234,36471,37026,370[illegible]37930,38023,38084,38158,39213,39284,39388,39758,40169,40303,40751,40932,41
014,41024,41080,41215,41248,41380,41382,414[illegible]41877,41898,41928,42413,42624,42875,43092,43114,43131,43152,43184,43239,43
556,43851,44226,44482,44557,44732,45109,45169,45496,45529,45585,45697,45835,45861,45869,45968,46011,46049,46600,47143,47227,47515,48355,48
468,48537,48574,48817,49090,49401,49488,49762,49819,50141,50583,51063,51302,51354,51501,51666,52046,52092,52097,52209,52218,52402,52718,52
817,52824,53006,53041,53278,53426,53659,53816,53877,53904,53911,54481,54529,54617,54879,54941,54984,54992,55124,55283,55387,56033,56412,56
439,56457,56594,56726,56887,56946,57003,57223,57346,57532,57659,57705,58708,58768,59143,59261,59350,59366,59633,59856,59919,59924,60184,60
229,60396,60476,60622,61400,61414,61465,61847,61876,61944,62323,62554,62637,62886,62898,63000,63116,63389,63415,63428,63442,63577,63918,64
305,64570,64665,64677,64774,65169,65336,65422,65454,65522,65602,65725,65913,66495,66535,67327,67986,68402,68460,68711,68721,68863,69038,69
051,69079,69361,70519,70528,70540,70764,70872,71018,71107,71195,71209,71259,71280,71498,71659,71665,71760,71830,71849,72133,72225,72352,72
664,72928,73298,73317,73414,73516,73563,73829,73928,74012,74125,74242,74279,74581,74594,75060,75107,75441,75662,75793,75818,75826,76221,76
746,77061,77088,77176,77322,77495,77811,77821,78123,78128,78173,78439,78471,78483,79070,79374,79499,79723,79749,79816,80052,80080,80160,80
177,80405,80903,81070,81190,81256,81417,81680,81795,82003,82148,82172,82199,82615,82669,82785,82820,82965,83017,83391,83412,84172,84323,84
577,84724,84845,84971,85003,85094,85130,85302,85517,85650,85702,85792,86151,86183,86189,86272,86321,86646,86820,86912,86918,87257,87417,87
525,87563,87773,87815,87850,87966,88002,88722,88787,88855,88925,89200,89226,89263,89832,89942,89958,90031,90281,90553,90667,90713,90891,91
493,91522,91993,92008,92333,93090,93740,93761,93920,93998,94224,94269,94456,94484,94669,94754,95282,95710,96269,96353,96478,98108,98154,98
570,98961,99780,99811,99878,99885,100177,100442,101699,101788,101803,101829,102130,102204,102261,102299,102437,102492,102904,102998,103166
,103551,103697,103928,104328,104540,104703,104853,105084,105196,105254,105310,105316,105488,105885,106019,106603,107302,107333,107378,1077
14,107781,108147,108173,108193,108196,108199,108491,108733,109007,109326,109450,109992,110118,110284,111096,111122,111199,111208,111537,11
2290,112345,112565,112678,112693,113005,113113,113164,113212,113268,113288,113522,113751,113837,113849,114004,114136,114227,114365,114445,
115502,115508,115527,115564,115676,115948,116096,116170,116266,116486,116543,117033,117991,118044,118469,118957,119028,119151,119246,11956
3,119777,120098,120144,120154,121120,121417,121426,121463,121470,121507,121522,121599,121820,121948,121955,121992,122004,122137,122239,122
266,122272,122300,123362,124006,124211,124448,124830,125169,125225,125281,125312,125330,125410,125532,126099,126186,126245,126524,126538,1
26570,126666,126819,126851,126864,126895,126908,[illegible]127090,127161,127252,127266,127305,128449,128686,128706,128842,128852,128883
,128968,129120,129299,129453,129577,129678,12991[illegible]97,130676,130753,130790,130829,131003,131076,131091,131130,131471,1315
01,131566,131597,131771,131818,132096,132767,133[illegible]4262,134370,134412,134426,134626,134772,134795,134798,134880,134938,13
5108,135174,135407,135481,135498,135532,136110,1[illegible]36590,136902,136954,137001,137122,137362,137427,137842,137862,138010,
138194,138347,138365,138729,138784,139110,139177[illegible]1,139981,140009,140234,140423,140448,140685,140767,140902,140955,14102
3,141096,141269,142599,142960,142999,143901,1439[illegible]120,144298,144447,144547,145099,145192,145326,145946,145980,146235,146
250,146573,146638,146695,146801,146948,146956,14[illegible]47403,147514,147846,148098,148480,148732,148805,148958,149163,149205,1
49331,149386,149553,149617,149950,150063,150357,1506[illegible]1,150737,150856,150975,151219,151447,151669,151691,151975,151999,152010,152075,152297
,152354,152524,152862,152871,152906,153630,153962,153979,153998,154144,154180,154311,154456,154465,154486,154760,154991,155841,155854,1563
07,156854,156874,157064,157282,157474,157507,157550
2       mansfield park.txt      159,545 861     5.40    0.957   159,545 words in text file plot positions:
273,337,444,510,609,700,732,775,814,[illegible]977,1020,1148,1653,1700,2309,2320,2479,2566,2610,3548,3680,3918,3957,4120,4308,4317,4331,4721,
4864,4985,5105,5276,5980,5992,6222[illegible]6654,6849,6961,7144,7522,7642,7652,8014,8280,8526,8577,8647,8787,8807,8845,8897,8913,9041,905
```

and you can then process those numbers in another program of your choice.

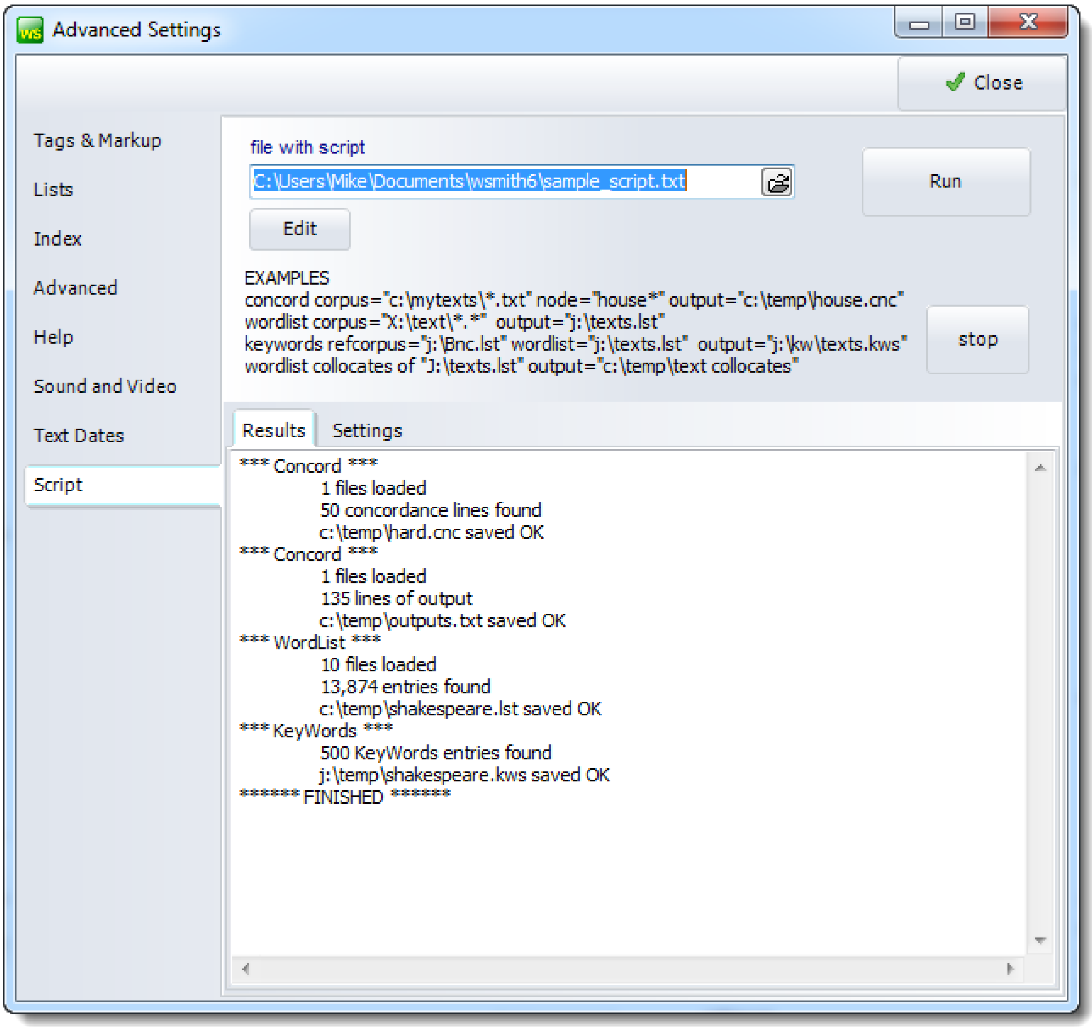
In the case of XML text (🌐), you get a little .HTM file and a large .XML file. Click on the .HTM file and you can see your data a page at a time, with buttons to jump forwards or back a page, as well as to the first and last pages of data. This accesses your .XML file to read the data itself.

See also: Excel Files in batch processing 39

5.28 scripting

Scripts

This option allows you to run a pre-prepared script. In the case below, **`sample_script.txt`** has requested two concordance operations, a word list, and a keywords analysis. The whole process happened without any intervention from the user, using the defaults in operation.



file with script: C:\Users\Mike\Documents\wsmith6\sample_script.txt

```
EXAMPLES
concord corpus="c:\mytexts\*.txt" node="house*" output="c:\temp\house.cnc"
wordlist corpus="X:\text\*.*"  output="j:\texts.lst"
keywords refcorpus="j:\Bnc.lst" wordlist="j:\texts.lst"  output="j:\kw\texts.kws"
wordlist collocates of "J:\texts.lst" output="c:\temp\text collocates"
```

```
*** Concord ***
        1 files loaded
        50 concordance lines found
        c:\temp\hard.cnc saved OK
*** Concord ***
        1 files loaded
        135 lines of output
        c:\temp\outputs.txt saved OK
*** WordList ***
        10 files loaded
        13,874 entries found
        c:\temp\shakespeare.lst saved OK
*** KeyWords ***
        500 KeyWords entries found
        j:\temp\shakespeare.kws saved OK
******* FINISHED *******
```

The syntax is as suggested in the EXAMPLES visible above. (There is a **`sample_script.txt`** file in your Documents\wsmith6 folder). First the tool required, then the necessary parameters, each surrounded by double quotes, in any order.

`concord corpus="x:\text\dickens\hard_times.txt" node="hard" output="c:\temp\hard.cnc"`

made a concordance of the hard_times.txt text file looking for the search-word hard and saved results in c:\temp\hard.cnc

`concord corpus="x:\text\dickens\hard_times.txt" node="c:\temp\sws.txt" output="c:\temp\outputs.txt" 1_at_a_time="true"`

made a concordance of the same text file looking for each search-word in the **`sws.txt`** file, counted the number of hits and saved results in **c:\temp\outputs.txt**

```
wordlist corpus="x:\text\shakes\oll\txt\tragedies\*.txt" output="c:\temp
\shakespeare.lst"
```

made a word list of all the .txt text files in a folder of Shakespeare tragedies (not including sub-folders) and saved it.

```
keywords refcorpus="j:\temp\BNC.lst" wordlist="c:\temp\shakespeare.lst"
output="j:\temp\shakespeare.kws"
```

made a key words list of that word list compared with a BNC word list and saved it.

Two additional optional parameters not visible there are:
1_at_a_time="true" and **TXT_format="true"**.

If **TXT_format** is true, a Concord file will contain only the concordance lines, a KeyWords file only the key words and their frequencies, and a WordList file only the words and their frequencies.

If **1_at_a_time** is true, a word-list will export separate results text file by text file.
If **1_at_a_time** is true, Concord will read search words from a text file and save summary results:

```
concord corpus="x:\text\dickens\hard_times.txt" node="c:\temp\sws.txt"
output="c:\temp\outputs.txt" 1_at_a_time="true"
```

produced this in **c:\temp\outputs.txt**:

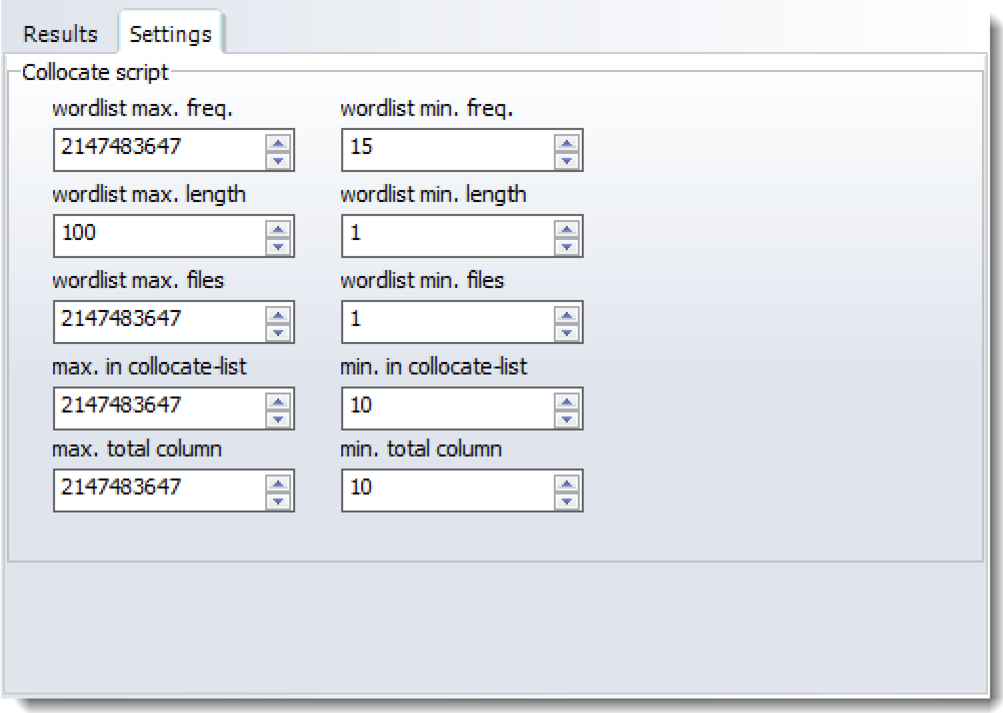
```
x:\text\dickens\hard_times.txt
hard      50
soft      3
mean      54
empty     9
fred      0
book      13
north*    4
south*    2
```

collocate scripts

It is also possible to run a script requesting the collocates of each word in a word-list. This syntax

```
wordlist collocates of "c:\temp\shakespeare.lst" output="c:\temp
\shakespeare\collocates"
```

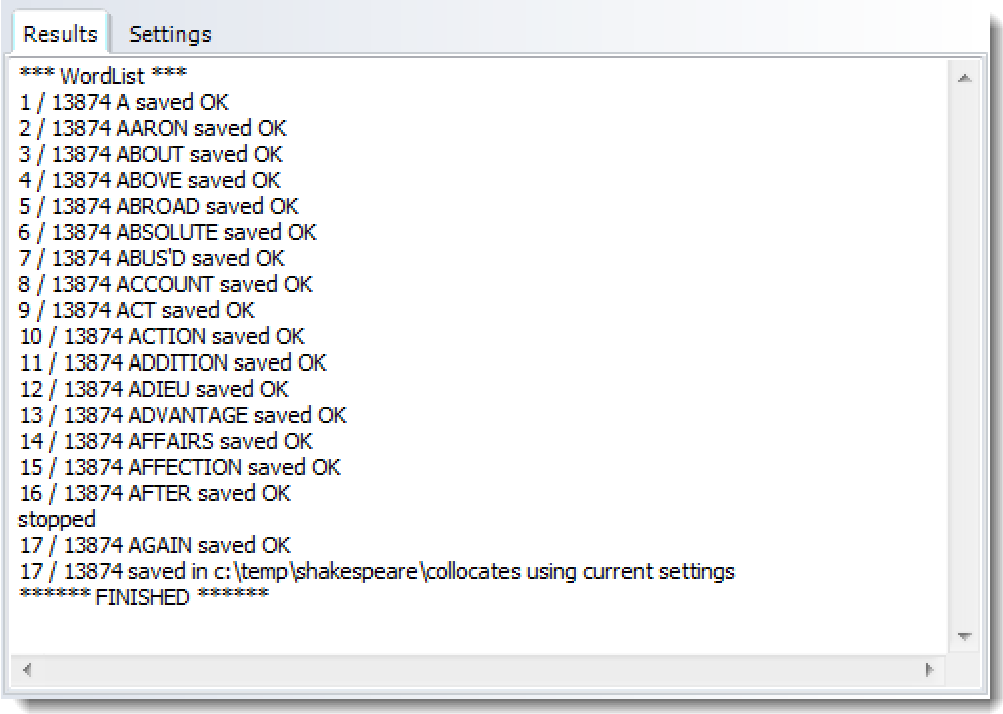
tells WordSmith to compute the collocates of each word in the **shakespeare.lst** word-list, and save results as plain text files, one per word, in the **c:\temp\shakespeare\collocates** folder. The texts to be processed are the same text files used when the word list was created (and must still be present on disk to work, of course). Settings affecting the process are shown below. The first 6 have to do with the words from the word-list, and the *min. in collocate-list* refers to how many collocates of each word-list word are needed (here 10) for processing to be reported. *Min. total column* refers to the number in the *total* column of a collocation display.



Collocate script

wordlist max. freq. 2147483647	wordlist min. freq. 15
wordlist max. length 100	wordlist min. length 1
wordlist max. files 2147483647	wordlist min. files 1
max. in collocate-list 2147483647	min. in collocate-list 10
max. total column 2147483647	min. total column 10

Results look like this:



```
*** WordList ***
1 / 13874 A saved OK
2 / 13874 AARON saved OK
3 / 13874 ABOUT saved OK
4 / 13874 ABOVE saved OK
5 / 13874 ABROAD saved OK
6 / 13874 ABSOLUTE saved OK
7 / 13874 ABUS'D saved OK
8 / 13874 ACCOUNT saved OK
9 / 13874 ACT saved OK
10 / 13874 ACTION saved OK
11 / 13874 ADDITION saved OK
12 / 13874 ADIEU saved OK
13 / 13874 ADVANTAGE saved OK
14 / 13874 AFFAIRS saved OK
15 / 13874 AFFECTION saved OK
16 / 13874 AFTER saved OK
stopped
17 / 13874 AGAIN saved OK
17 / 13874 saved in c:\temp\shakespeare\collocates using current settings
****** FINISHED ******
```

Here they're incomplete because I pressed the Stop button.

Each of these lists has the collocates output much as in a collocates display 181, but with the relationships also computed.

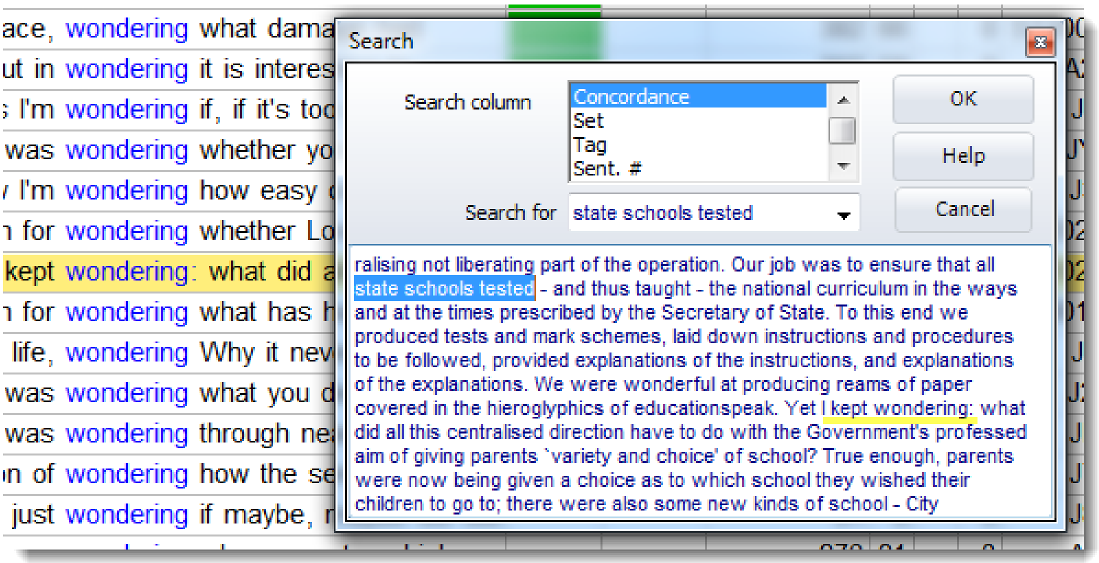
The process only saves results where the settings shown above are met and where the relationships also meet the requirements as in the WSConcgram 395 settings.

See also : [drag and drop](427)

5.29 searching

5.29.1 search for word or part of word

All lists allow you to search for a word or part of one, or a number. The search operates on the *current column* of data, though you can change the choice as in this screenshot, where Concordance is selected.



```
ace, wondering what dama
ut in wondering it is interes
; I'm wondering if, if it's too
was wondering whether yo
/ I'm wondering how easy
ı for wondering whether Lo
kept wondering: what did a
ı for wondering what has h
life, wondering Why it nev
was wondering what you d
was wondering through ne
ın of wondering how the se
just wondering if maybe, r
```

Search for: state schools tested

ralising not liberating part of the operation. Our job was to ensure that all state schools tested - and thus taught - the national curriculum in the ways and at the times prescribed by the Secretary of State. To this end we produced tests and mark schemes, laid down instructions and procedures to be followed, provided explanations of the instructions, and explanations of the explanations. We were wonderful at producing reams of paper covered in the hieroglyphics of educationspeak. Yet I kept wondering: what did all this centralised direction have to do with the Government's professed aim of giving parents `variety and choice' of school? True enough, parents were now being given a choice as to which school they wished their children to go to; there were also some new kinds of school - City

The syntax is as in Concord. In the case of a concordance line, the search operates over the whole context so far saved or retrieved. So although **`kept wondering`** is visible in the context (highlighted to show you where) the search has found the phrase **`state schools tested`** about 80 words before the search word **`wondering`**.

To search again, press OK again....

Whole word – or bung in an asterisk

The syntax is as in Concord, so by default a whole word search. To search for a suffix or prefix, use the asterisk. Thus **`*ed`** will find any entry ending in **`ed`**; **`un*`** will find any entry starting with **`un.`** **`*book*`** will find any entry with **book** in it (**`book, textbook, booked.`**)

See also: [Searching by Typing](109), [Search & Replace](110).

5.29.2 search by typing

Whenever a column of display is organised alphabetically, you can quickly find a word by typing. As you type, **WordSmith** will get nearer. If you've typed in the first five letters and **WordSmith** has found a match, there'll be a beep, and the edit window will close. You should be able to see the word you want by now.

See also: Edit v. Type-in mode 428, Searching for a word or part of one 109, Search & Replace 110, Editing 72, WordList sorting 315

5.29.3 search & replace

Some lists, such as lists of filenames 113, allow for searching and replacing.

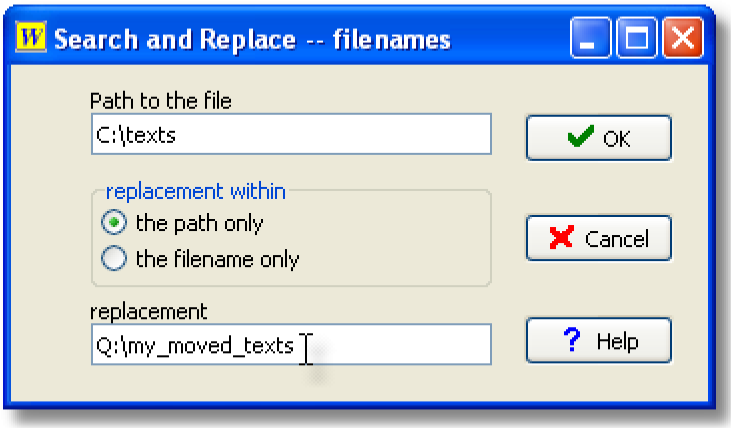
The point of it

If your text data has been moved from one PC to another, or one drive to another, it will be necessary to edit all the filenames if WordSmith ever needs to get at the source texts, such as when computing a concordance from a word list 234.)

Search & Replace for filenames

If you are replacing a filename you will see something like this. We distinguish between the path and the file's individual name, so that for a case like **`C:\texts\BNC\spoken\s conv\KC2.txt`** the filename is **`KC2.txt`** and the path to it is **`C:\texts\BNC\spoken\s conv`**.

To correct the path to the file, e.g. if you've moved your BNC texts to drive **`Q:\my_moved_texts`** you might simply replace as shown here



Search and Replace -- filenames

Path to the file: C:\texts

replacement within
● the path only
○ the filename only

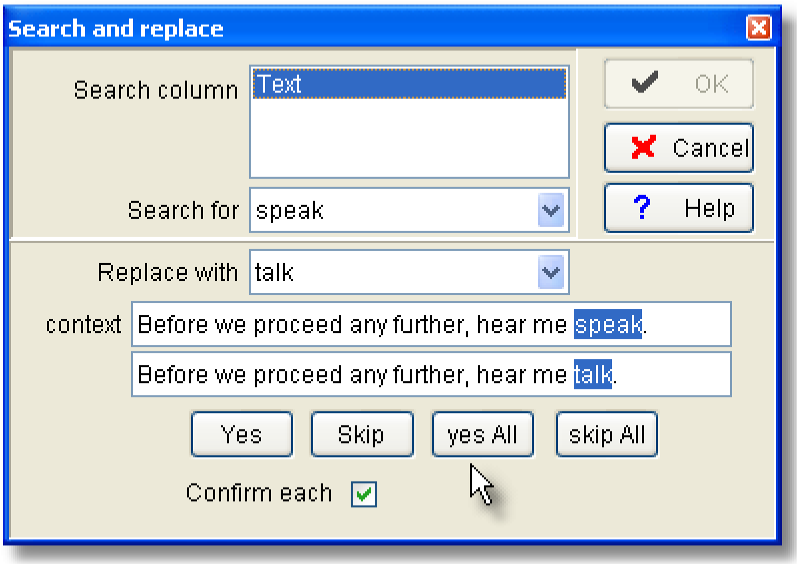
replacement: Q:\my_moved_texts

and all the filenames which contain **`c:\texts`** will get **`Q:\my_moved_texts`** e.g. **`C:\texts\BNC\spoken\s conv\KC2.txt`** will become **`Q:\my_moved_texts\BNC\spoken\s conv\KC2.txt.`**

To rename a filename only, change the radio buttons in the middle of the window and the search and replace operation will ignore the path but replace within the filename only.

Search & Replace for other data

In this case the search & replace isn't of filenames but in the case below in *Viewer and Text Aligner*, of the actual text. Like a search 109 operation, the search operates on the *current column* of data.



Search and replace

Search column: Text

Search for: speak

Replace with: talk

context: Before we proceed any further, hear me speak.

Before we proceed any further, hear me talk.

Confirm each ☑

The context line shows what has been found.

The line below shows what will happen if you agree to the change.

Yes: make 1 change (the highlighted one), then search for the next one

Skip: leave this one unchanged, search for the next one

Yes All: change without any check

Skip All: stop searching...

Whole word – or bung in an asterisk

The syntax is as in Concord, so by default a whole word search. To search for a suffix or prefix, use the asterisk. Thus **`*ed`** will find any entry ending in **`ed`**; **`un*`** will find any entry starting with **`un.`** **`*book*`** will find any entry with **`book`** in it (**`book, textbook, booked.`**)

Word lists can be sorted by suffix: see WordList sorting 315.

See also: Searching by Typing 109, Searching with F12 109, Accented Characters & Symbols 420.

5.30 selecting or marking entries

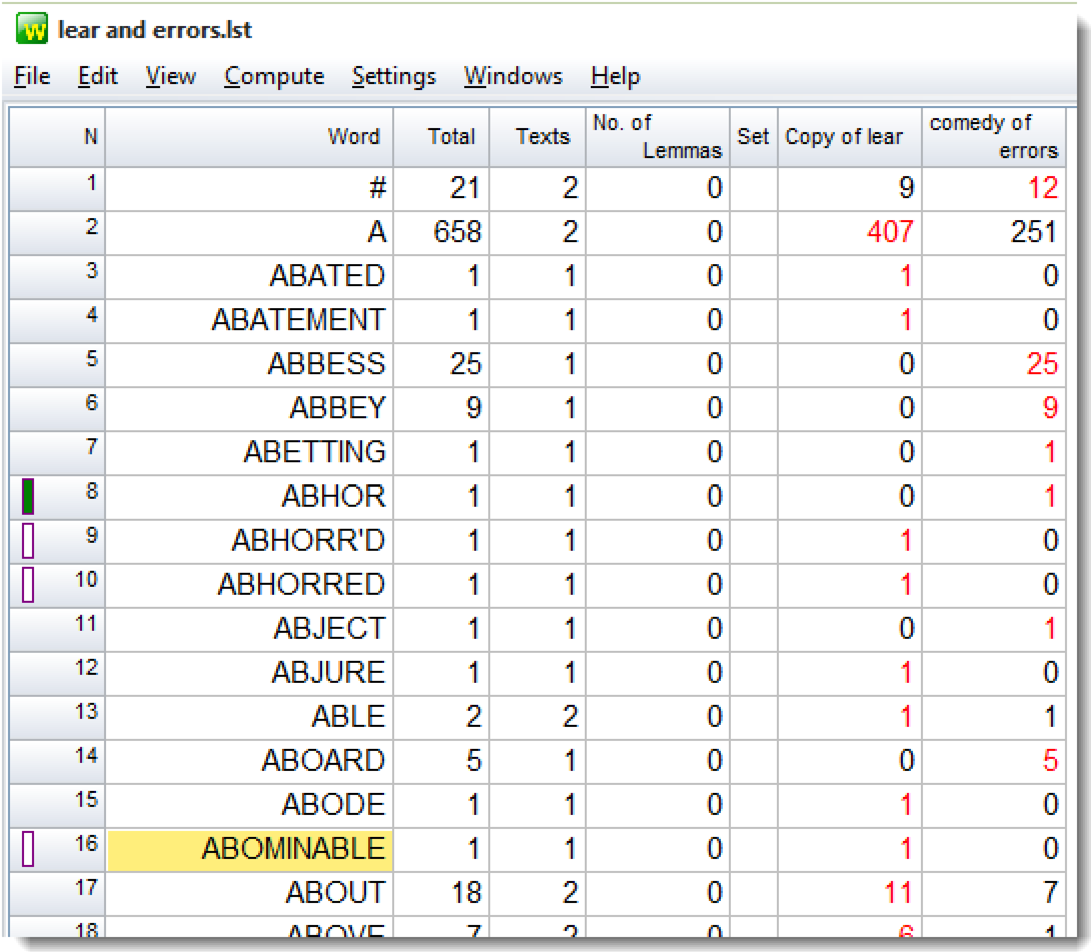
5.30.1 selecting multiple entries

It can be necessary to select non-adjacent entries e.g. for Joining 270 (lemmatisation).

How to do it

To select more than one entry in a word-list, concordance, key word list etc, hold down Control first, and click in the number column at the left edge.

To select various entries in this detailed consistency list, I held down the *Ctrl* key and clicked at the numbers 8, 9, 10 and 16.

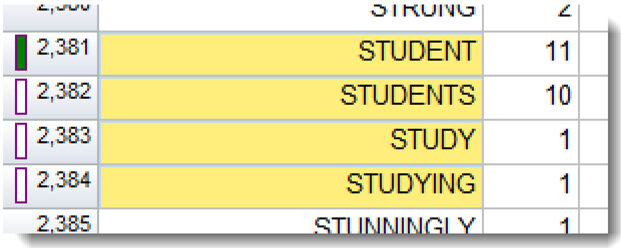


N	Word	Total	Texts	No. of Lemmas	Set	Copy of lear	comedy of errors
1	#	21	2	0		9	12
2	A	658	2	0		407	251
3	ABATED	1	1	0		1	0
4	ABATEMENT	1	1	0		1	0
5	ABBESS	25	1	0		0	25
6	ABBEY	9	1	0		0	9
7	ABETTING	1	1	0		0	1
🟩 8	ABHOR	1	1	0		0	1
□ 9	ABHORR'D	1	1	0		1	0
□ 10	ABHORRED	1	1	0		1	0
11	ABJECT	1	1	0		0	1
12	ABJURE	1	1	0		1	0
13	ABLE	2	2	0		1	1
14	ABOARD	5	1	0		0	5
15	ABODE	1	1	0		1	0
□ 16	ABOMINABLE	1	1	0		1	0
17	ABOUT	18	2	0		11	7
18	ABOVE	7	2	0		6	1

Alternatively, to mark entries you can choose *Edit* | Mark (F5) 112 in the menu.

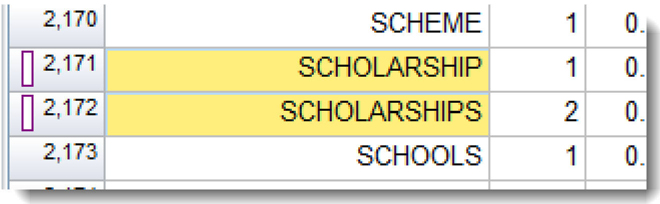
5.30.2 marking entries

Non-adjacent entries can be marked by clicking the word and pressing Alt+F5. The first one marked will get a green mark in the margin and subsequent ones will get white marks.



2,380	STRUNG	2
▮ 2,381	STUDENT	11
□ 2,382	STUDENTS	10
□ 2,383	STUDY	1
□ 2,384	STUDYING	1
2,385	STUNNINGLY	1

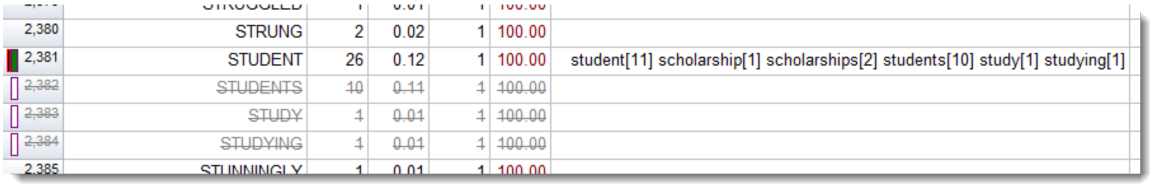
After marking these, I added **`scholarship(s)`** by selecting two more words then pressing Alt+F5 again.



2,170	SCHEME	1	0.
□ 2,171	SCHOLARSHIP	1	0.
□ 2,172	SCHOLARSHIPS	2	0.
2,173	SCHOOLS	1	0.

To undo a specific entry, press Alt+F5 again.
To un-mark all entries, use Shift+Alt+F5.

To lemmatise all marked entries, press F4 after marking.



2,379	STRUGGLED	1	0.01	1	100.00	
2,380	STRUNG	2	0.02	1	100.00	
▮ 2,381	STUDENT	26	0.12	1	100.00	student[11] scholarship[1] scholarships[2] students[10] study[1] studying[1]
□ ~~2,382~~	~~STUDENTS~~	~~10~~	~~0.11~~	~~1~~	~~100.00~~	
□ ~~2,383~~	~~STUDY~~	~~1~~	~~0.01~~	~~1~~	~~100.00~~	
□ ~~2,384~~	~~STUDYING~~	~~1~~	~~0.01~~	~~1~~	~~100.00~~	
2,385	STUNNINGLY	1	0.01	1	100.00	

5.31 filenames tab

This tab shows the text file name(s) from which your current data comes. You can edit these names if necessary (e.g. if the text files have been moved or renamed.) To do so, choose Replace ( ).
Afterwards, if you save the results 101, the information will be permanently recorded.

See also: finding source files 430.

5.32 settings

5.32.1 save defaults

Settings can be altered by choosing *Colour Settings* in the WordSmith Tools Controller 4.
Any setting menu item in any Tool gives you access to these:

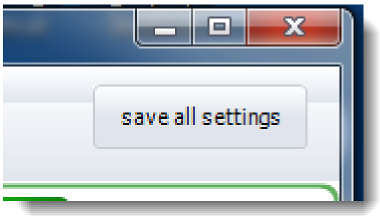
General, Folders, Colours, Languages, Tags, Lists, Concord, KeyWords, WordList, Index, Advanced, WSConcgram

These tabs allow you to choose settings which affect one or more of the Tools.

colours 60	customise the default colours
folders 431	set WordSmith so it "knows" which folders you usually use
languages 124	character set 419, treatment of hyphens 435 & numbers, default file extension
general	restore last file 447, printing 80
tags 132	tags to ignore, tag file, tag file autoloading, custom tagsets 133
stop lists 120	for Concord, KeyWords and Wordlist
matching	files 92 to match up, or lemma files 270 to mark lemmas in a word list, etc.
Concord	number of entries, sort system, collocation horizons 180
KeyWords	procedure 245, max. p value 235, database & associate minimum frequencies, reference corpus 447 filename
WordList	word length & frequencies, type/token # 303, cluster 448 settings
Index 276	making a word-list index
Advanced 31	advanced settings
WSConcgram	for the concgram 12 utility

permanent settings and wordsmith6.ini file

You can save your settings with a button at the top of the Controller



Or by editing the **wordsmith6.ini** file, installed when you installed WordSmith Tools. This specifies all the settings which you regularly use for all the suite of programs, such as your text and results folders 431, screen colours 60, fonts 78, the default columns 87 to be shown in a concordance, etc.

You can restore the defaults 115.

show help file

In the general tab of Main Settings you will see a checkbox called "show help file". If checked, this will always show this help file every time WordSmith starts up. The point of this is for users who only use the software occasionally, e.g. in a site licence installation.

sayings

Using Notepad, you can edit **Documents\wsmith6\sayings.txt**, which holds sayings that appear in the main Controller 4 window, if you don't like the sayings or want to add some more.

site licence and CD-ROM defaults

If you're running WordSmith straight from a CD-ROM, your defaults cannot be saved on it as it's read-only; Windows will find a suitable place for **`wordsmith6.ini`**, usually off the root folder of **`My Documents`**.

The first time you use WordSmith, you will be prompted to choose appropriate Folders 431, Text 124 Characteristics, Tag 132 details etc. and *Save All Settings* for future use. You can change settings and save them as often as you like.

Similarly, on a network you will usually not be allowed to change defaults permanently, as this would affect other users. Your network administrator should have installed the program so that you have your own copy of **`wordsmith6.ini`**, where it may be both read and altered. If WordSmith Tools finds a copy of **`wordsmith6.ini`** in that folder it will be able to use your personal preferences.

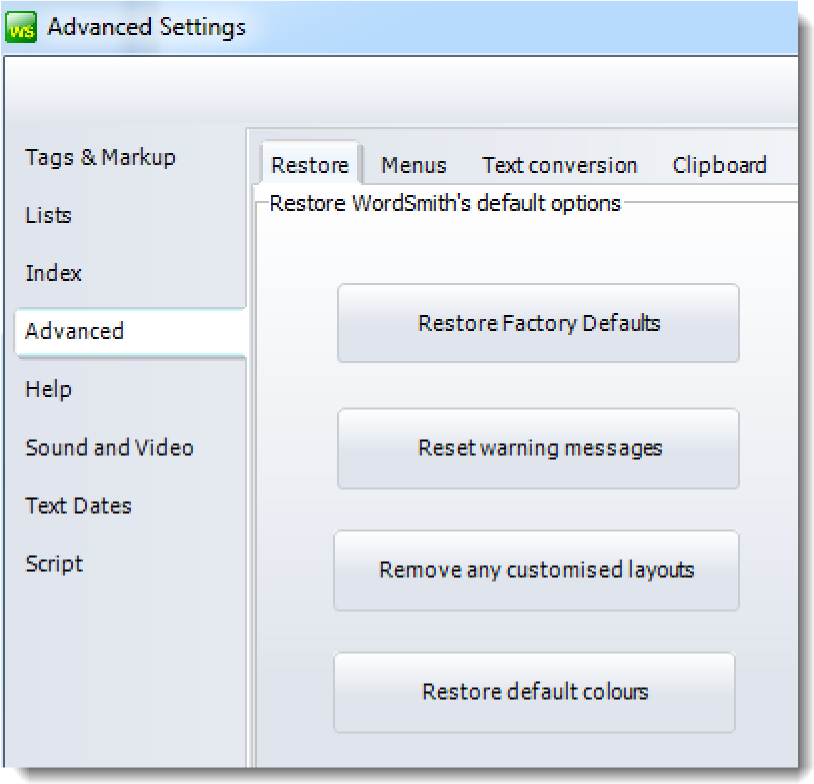
5.32.2 restoring settings

How to find it

Settings can be restored to default settings by choosing *Main settings* | *Advanced Settings* | *Advanced* | *Restore*.

The point of it...

You may have changed settings and cannot recall how to undo them....



Factory Defaults

Restores your `wordsmith6.ini` file to factory condition. Re-starts WordSmith with all relevant boxes filled in accordingly.

Warning Messages

Removes any of the messages you have received and where you've ticked the "never show this again" box.

Customised Layouts

Removes any of the layouts 87 you have saved for the various type of data WordSmith shows.

Colours

Re-sets colours 60 to the factory defaults.

5.33 source text(s)

The point of it...

The aim is to be able to see the whole text file your data came from, with some relevant words highlighted.

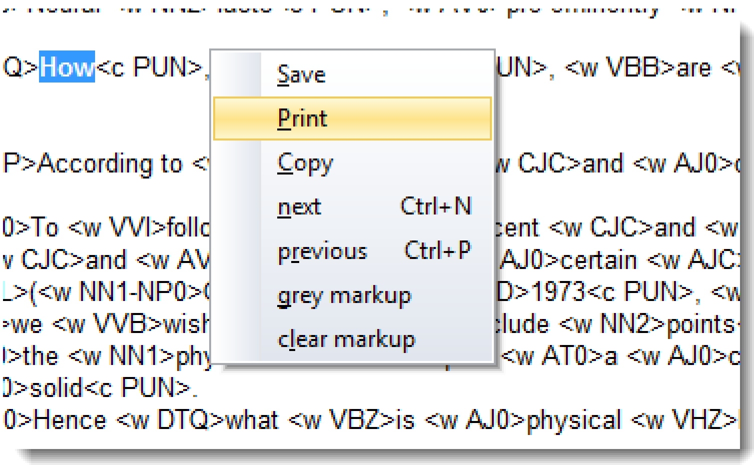
How to do it

The **Concord** and **KeyWords** tools both have areas which can show the source texts which your data was produced from, visible by choosing the *source texts* tab, if your texts are still where they were when the data analysis was done. (If they have been moved you can try editing the *Filenames* 110 data to correct this.)

In Concord, you need to double-click the relevant concordance line to get the source text to show. In each case the relevant key or search words will be highlighted if possible.

In KeyWords you'll see the source text in the source texts tab space, or if there are various source texts listed in a special window (shown below).

Menu options (right-click to see these)



Copy, Print, Save

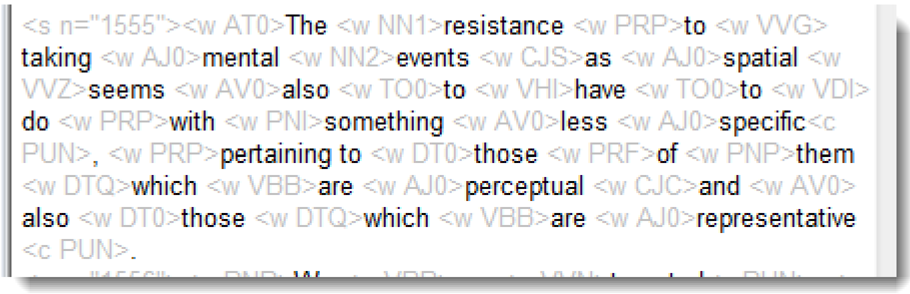
As their names suggest these menu items let you copy, save or print any text you've selected or the whole text. For saving you will get a chance to decide whether as plain text or as Rich Text Format (.RTF) preserving font and colour information.

Next, Previous

These jump you through the text one highlighted word at a time. You should see how many highlighted there are in the status bar.

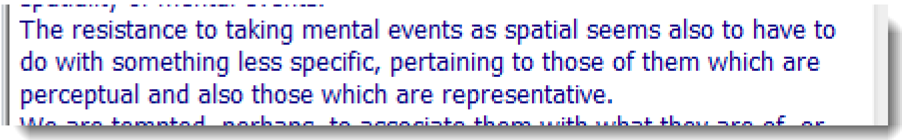
Grey markup, Clear markup, Restore markup

Grey markup lets you grey out all < > sections.



<s n="1555"><w AT0>The <w NN1>resistance <w PRP>to <w VVG>taking <w AJ0>mental <w NN2>events <w CJS>as <w AJ0>spatial <w VVZ>seems <w AV0>also <w TO0>to <w VHI>have <w TO0>to <w VDI>do <w PRP>with <w PNI>something <w AV0>less <w AJ0>specific<c PUN>, <w PRP>pertaining to <w DT0>those <w PRF>of <w PNP>them <w DTQ>which <w VBB>are <w AJ0>perceptual <w CJC>and <w AV0>also <w DT0>those <w DTQ>which <w VBB>are <w AJ0>representative <c PUN>.

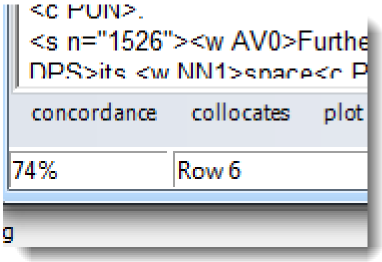
Clear markup simply cuts the tags out.



spatiality of mental events.
The resistance to taking mental events as spatial seems also to have to do with something less specific, pertaining to those of them which are perceptual and also those which are representative.
We are tempted, perhaps, to associate them with what they are of, or

Once you have cut out markup, the *clear markup* option changes to *restore markup* (needed if Concord is to jump to the correct location).

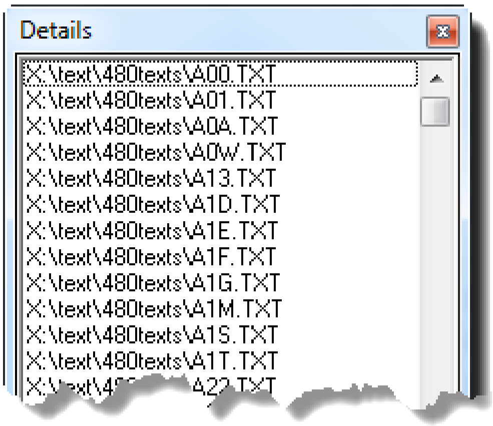
Greying out mark-up is quite slow if the text is extensive. This shot shows its progress:



<c PUN>.
<s n="1526"><w AV0>Furthe
DPS>its <w NN1>space<c P

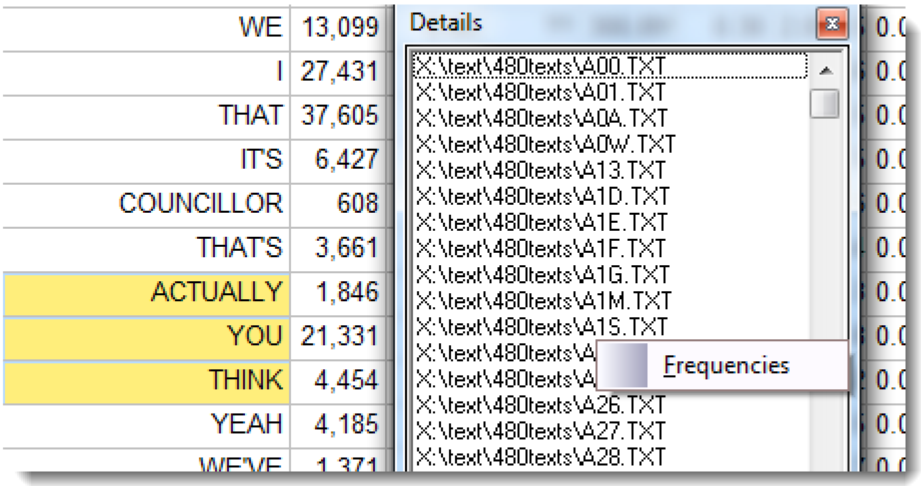
Double-clicking the status bar gives you a chance to stop the process.

KeyWords List of Source Texts



X:\text\480texts\A00.TXT
X:\text\480texts\A01.TXT
X:\text\480texts\A0A.TXT
X:\text\480texts\A0W.TXT
X:\text\480texts\A13.TXT
X:\text\480texts\A1D.TXT
X:\text\480texts\A1E.TXT
X:\text\480texts\A1F.TXT
X:\text\480texts\A1G.TXT
X:\text\480texts\A1M.TXT
X:\text\480texts\A1S.TXT
X:\text\480texts\A1T.TXT
X:\text\480texts\A22.TXT

If you right-click this window, you get a chance to see which texts contain which key words

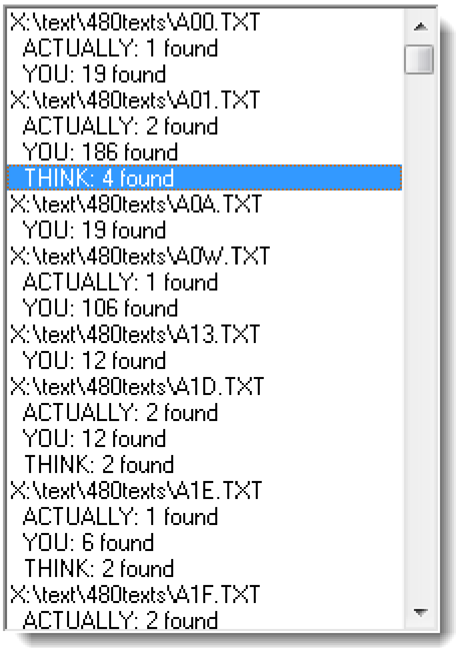


WE	13,099	0.(
I	27,431	0.(
THAT	37,605	0.(
IT'S	6,427	0.(
COUNCILLOR	608	0.(
THAT'S	3,661	0.(
ACTUALLY	1,846	0.(
YOU	21,331	0.(
THINK	4,454	0.(
YEAH	4,185	0.(
WE'VE	1,371	0.(

Details

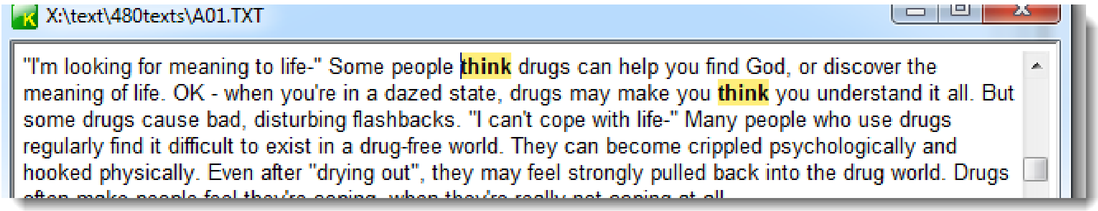
```
X:\text\480texts\A00.TXT
X:\text\480texts\A01.TXT
X:\text\480texts\A0A.TXT
X:\text\480texts\A0W.TXT
X:\text\480texts\A13.TXT
X:\text\480texts\A1D.TXT
X:\text\480texts\A1E.TXT
X:\text\480texts\A1F.TXT
X:\text\480texts\A1G.TXT
X:\text\480texts\A1M.TXT
X:\text\480texts\A1S.TXT
X:\text\480texts\A
X:\text\480texts\A
X:\text\480texts\A26.TXT
X:\text\480texts\A27.TXT
X:\text\480texts\A28.TXT
```

by clicking *Frequencies*, giving results like this:



```
X:\text\480texts\A00.TXT
  ACTUALLY: 1 found
  YOU: 19 found
X:\text\480texts\A01.TXT
  ACTUALLY: 2 found
  YOU: 186 found
  THINK: 4 found
X:\text\480texts\A0A.TXT
  YOU: 19 found
X:\text\480texts\A0W.TXT
  ACTUALLY: 1 found
  YOU: 106 found
X:\text\480texts\A13.TXT
  YOU: 12 found
X:\text\480texts\A1D.TXT
  ACTUALLY: 2 found
  YOU: 12 found
  THINK: 2 found
X:\text\480texts\A1E.TXT
  ACTUALLY: 1 found
  YOU: 6 found
  THINK: 2 found
X:\text\480texts\A1F.TXT
  ACTUALLY: 2 found
```

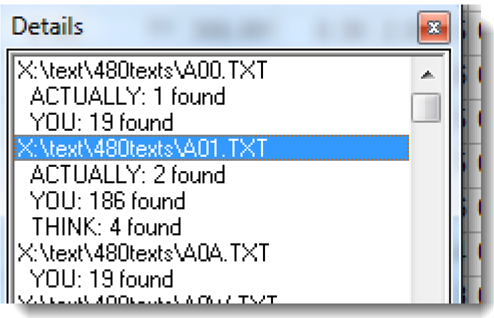
and if you double-click a highlighted word (**`THINK`** in the example above), you will be shown the source text (**`A01.txt`**) with that word highlighted.



X:\text\480texts\A01.TXT

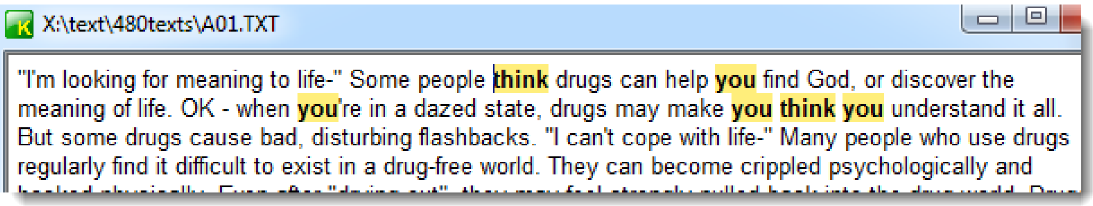
"I'm looking for meaning to life-" Some people think drugs can help you find God, or discover the meaning of life. OK - when you're in a dazed state, drugs may make you think you understand it all. But some drugs cause bad, disturbing flashbacks. "I can't cope with life-" Many people who use drugs regularly find it difficult to exist in a drug-free world. They can become crippled psychologically and hooked physically. Even after "drying out", they may feel strongly pulled back into the drug world. Drugs often make people feel they're coping, when they're really not coping at all.

If you simply click the file-name



X:\text\480texts\A00.TXT
ACTUALLY: 1 found
YOU: 19 found
X:\text\480texts\A01.TXT
ACTUALLY: 2 found
YOU: 186 found
THINK: 4 found
X:\text\480texts\A0A.TXT
YOU: 19 found

you get to see the text with all its key words highlighted.



"I'm looking for meaning to life-" Some people **think** drugs can help **you** find God, or discover the meaning of life. OK - when **you**'re in a dazed state, drugs may make **you** **think** **you** understand it all. But some drugs cause bad, disturbing flashbacks. "I can't cope with life-" Many people who use drugs regularly find it difficult to exist in a drug-free world. They can become crippled psychologically and

5.34 stop lists

Stop lists are lists of words which you don't want to include in analysis. For example you might want to make a word list or analyse key words excluding common function words like *the, of, was, is, it*.

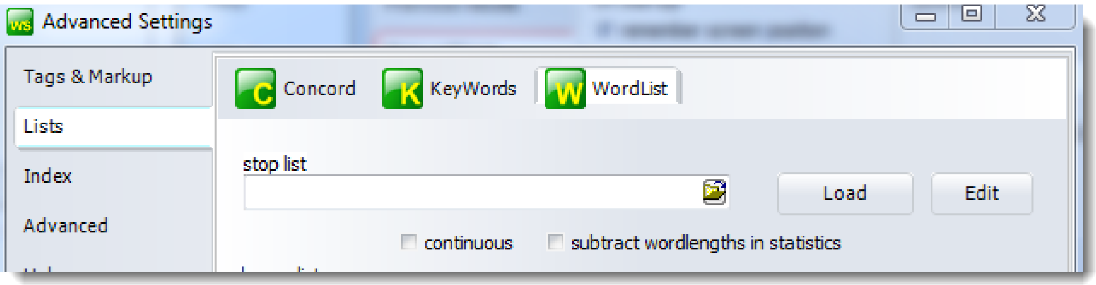
To use stop lists, you first prepare a file, using **Notepad** or any plain text word processor, which specifies all the words you wish to ignore. Separate each word using commas, or else place each one on a new line. You can use capital letters or lower-case as you prefer. You can use a semi-colon for comment lines. There is no limit to the number of words. Stop lists do not use wildcards (match-lists 92 may).

There is a file called **stop_wl.stp** (in your **\wsmith6** folder) which you could use as a basis and save under a new name. You'll also find **basic_English_stoplist.stp** there, based on top frequency items in the BNC. Or just make your own in Notepad and save it with **.stp** as the file-extension. If that is difficult, rename the **.txt** as **.stp**.

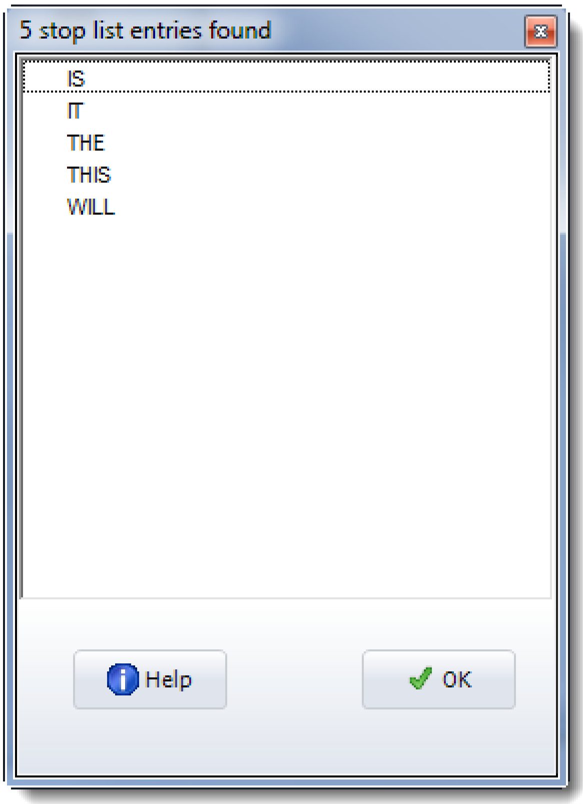
Example

```
; My stop list for test purposes.
THE,THIS,IS
IT
WILL
```

Then select *Stop List* in the menu to specify the stop list(s) you wish to use. Separate stop lists can be used for the **WordList, Concord** and **KeyWords** programs. If the stop list is *activated*, it is in effect: that is, the words in it will be stopped from being included in a word list. If you wish always to use the same stop list(s) you can specify them in **`wordsmith6.ini`** as defaults 113.



To choose your stop list, click the small yellow button in the screenshot, find the stop list file, then press *Load*. You will see how many entries were correctly found and be shown the first few of them.


5 stop list entries found

IS
IT
THE
THIS
WILL

With a stop list thus loaded, start a new word list. The words in your stop list should now not appear in the word list.

continuous

Normally, every word is read in while making the word list and stored in the computer's memory without checking whether it's the stop list. Eventually the set of words is checked in your stop list and omitted if it is present. That is much quicker. However, it means that for the most part, any statistics 298 are computed on the whole text, disregarding your stop list.

If you choose *continuous* the processing will slow down dramatically since as every word is read in while making the word list, it will be checked against the stop list and ignored if found. In other words, *every single case* of **THE** and **OF** and **IS** etc. will be looked at as the texts are read in and sought in your stop list. The effect will be to give you detailed statistics which ignore the words in the stop lists.

subtract wordlengths in statistics

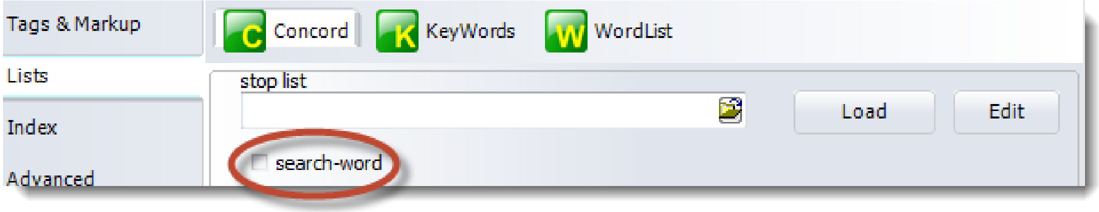
If you have not chosen continuous processing as explained above, you may want the statistics of your word list to attempt to deal in part with the stop list work done. With this choice, after the word list is computed, all the statistics concerning the number of types and tokens and 3-letter, 4-letter words etc. will be adjusted for the overall column (but not for the column for each single text) in your statistics 298.

See Match List 92 for a more detailed explanation, with screenshots.

Another method of making a stop list file is to use **WordList** on a large corpus of text, setting a high minimum frequency if you want only the high-frequency words. Then save it as a text file. Next, use the **Text Converter** to format it, using **stoplist.cod** as the Conversion file 362.

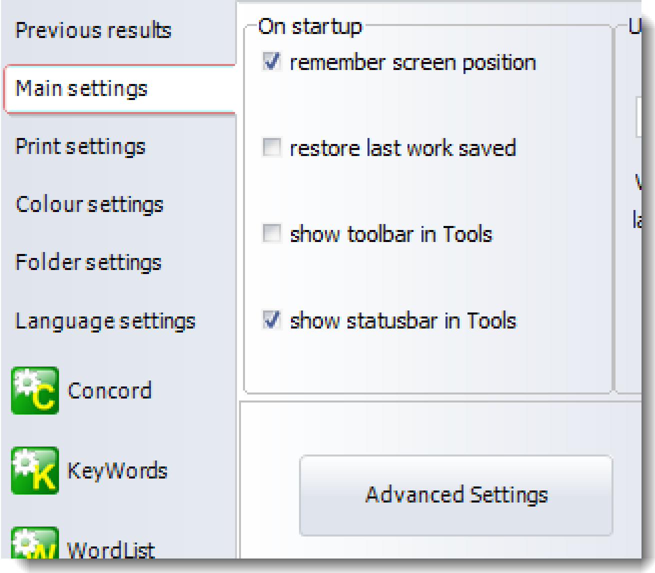
stop lists in Concord

In the case of Concord a stop list can do two jobs: first, it will cut the stop list words out as collocates. Additionally it can cut out any stop list words as search-word hits: for example if you concordance **beaut*** and **beautiful** is in your stop list, any concordance lines containing **beautiful** will get cut out (those containing **beauty** will remain). For this to be activated, make sure you check the search-word box in the settings.



Stop lists

... are accessed via an Advanced Settings button in the Controller



On startup

☑ remember screen position

☐ restore last work saved

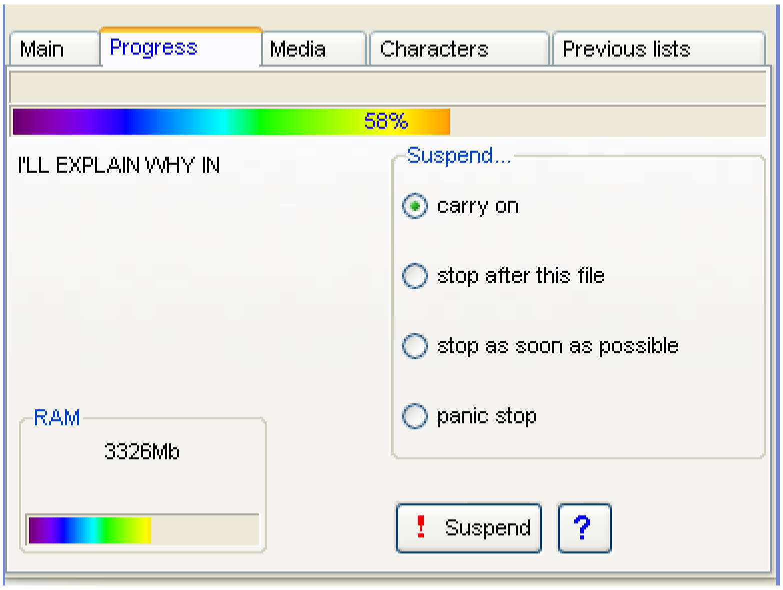
☐ show toolbar in Tools

☑ show statusbar in Tools

See also: Making a Tag File 141, Match List 92, Lemmatisation 274.

5.35 suspend processing

As WordSmith works its way through text files, or re-sorting data, you will see a progress window in the Controller with horizontal bars showing progress. If appropriate there'll be a *Suspend* button, too. Pressing this offers 4 choices:



I'LL EXPLAIN WHY IN

Suspend...

● carry on

○ stop after this file

○ stop as soon as possible

○ panic stop

RAM 3326Mb

carry on

... as if you had not interrupted anything

stop after this file

Finishing the file means that you can keep track of what has been done and what there wasn't time for. (How? By examining the filenames in the word list, concordance or whatever you have just been creating.)

stop as soon as possible

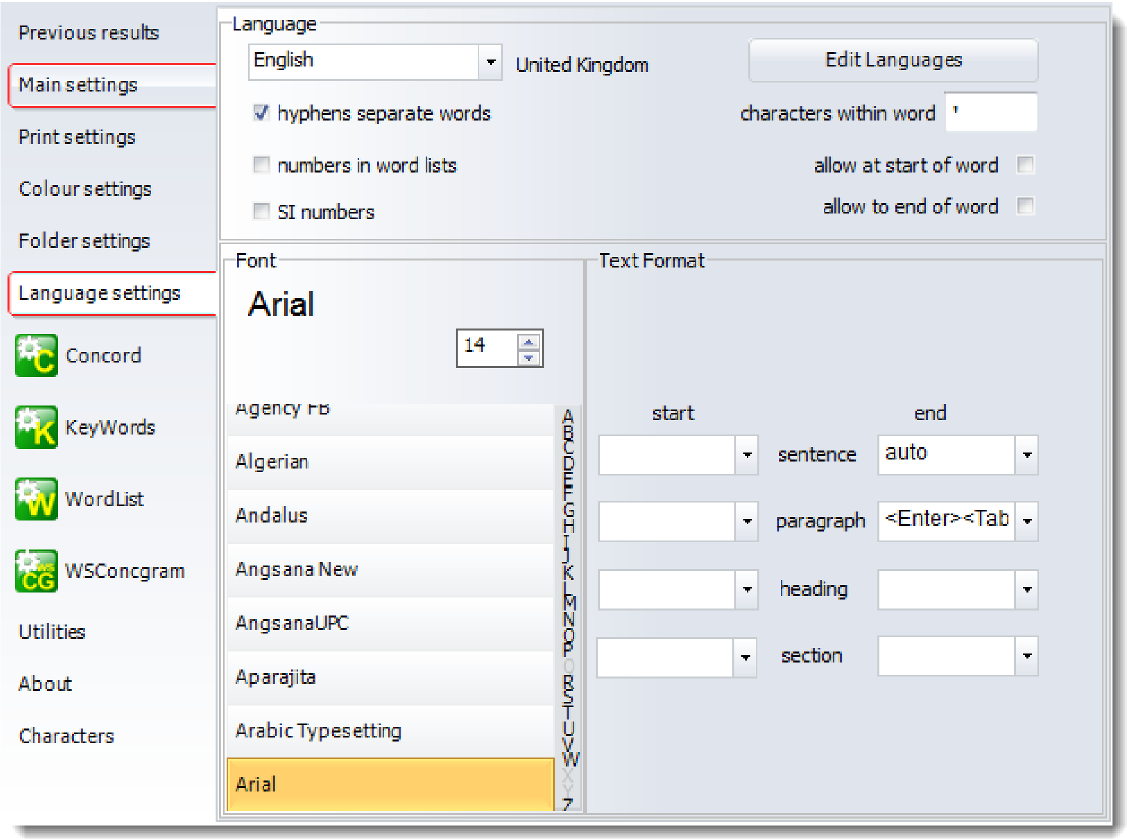
...useful if you're ploughing through massive CD-ROM files. WordSmith will stop processing the current file in the middle, but will retain any data it has got so far.

panic stop

The whole Tool (Concord or WordList, or whatever) will close down and some system resources memory 447 may be wasted. The Controller 4 will not be closed down.

Press *Suspend* again to effect your choice.

5.36 text and languages



Language: English, United Kingdom

☑ hyphens separate words

characters within word '

□ numbers in word lists

allow at start of word □

□ SI numbers

allow to end of word □

Font: Arial 14

sentence auto

paragraph <Enter><Tab

heading

section

These settings affect how WordSmith will handle your texts. At the top, you see boxes allowing you to choose the language family (eg. English) and sub-type (UK, Australia etc.). These choices are determined by the preferences you have previously set. That is, the expectation is that you only work with a few preferred languages, and you can set these preferences once and then forget about them. You do this by pressing the Edit Languages 7 button.

The choices below may differ for each language:

hyphens and numbers

You can also specify whether hyphens are to count as word separators. If the hyphen box is checked [X], **`self-access`** will be treated as two words.

Should numbers be included in a word-list as if they were ordinary words? If you leave this checkbox blank, words like $300, 50.3M or 10th will be ignored in word lists, key words, concordances etc. and replaced by a #. If you switch it on, they will be included. *SI numbers*: the International System of Units (SI) stipulates that "Spaces should be used as a thousands separator (1 000 000) in contrast to commas or periods (1,000,000 or 1.000.000) to reduce confusion resulting from the variation between these forms in different countries." So numbers like 1,234,567.89 would be written 1 234 567.89. If you wish WordSmith to recognise such forms as one number each, leave this box checked, otherwise such a form in text would be counted as three successive numbers (1, 234, and 567.89).

characters within word

WordSmith automatically includes as valid alphabetical symbols all those determined by the operating system as alphabetical for the language chosen. So, for English, A to Z and common accents such as **é**. For Arabic or Japanese, whatever characters Microsoft have determined count as alphabetic.

But you may wish to allow certain additional characters within a word. For example, in English, the apostrophe in **`father's`** is best included as a valid character as it will allow processing to deal with the whole word instead of cutting it off short. (If you change language to French you might not want apostrophes to be counted as acceptable mid-word characters.)

Examples:

'	(only apostrophes allowed in the middle of a word)
'%	(both apostrophes and percent symbols allowed in the middle of a word)
'_	(both apostrophes and underscore characters allowed in the middle of a word)
You can include up to 10.

If you want to allow **`fathers'`** too, check the *allow to end of word* box. If this is checked, any of these symbols will be allowed at either end of a word as long as the character isn't all by itself (as in **`"  '  "`**).

Plain Text/HTML/SGML

Your texts may be Plain Text in format: the default. If they are tagged 131 in HTML, SGML or XML 435 you should choose one of the options here. That way, the Tools can make optimum use of sentence, paragraph and heading mark-up.

(start & end of) headings

For the Tools to count headings, they need to know how to recognise the start and end of one. If your text is tagged 131 e.g. with **`<h1>`** and **`</h1>`**, type **`<h#>`** and **`</h#>`** in here. (# stands for any digit, ## for two, etc.) Whatever you type is case sensitive: **`</H#>`** is not the same as **`</h#>`**. (If you have HTML 435 text which is not consistent, using sometimes **`</h1>`** and sometimes **`</H1>`**, then use Text Converter 9 to make your texts consistent).

(start & end of) sections

If these boxes contain eg. **`<div#>`** and **`</div>`**, the Tools will treat identify sections. Again, whatever you type is case sensitive.

(start & end of) **sentences**

If this space contains the word **`auto`**, the Tools will treat sentences as defined 425 (ending with a full stop, question mark or exclamation mark, and followed by a capital letter), but if your text is tagged 131 e.g. with **`<s>`** and **`</s>`**, type those in here. Again, whatever you type is case sensitive.

(start & end of) **paragraphs**

For the Tools to recognise paragraphs, they need to know what constitutes a paragraph start and/or end, e.g. a sequence of two <Enter>s (where the original author pressed Enter twice) or an <Enter> followed by a <Tab>. For that you would type **`<Enter><Tab>`**. If your text is tagged 131 e.g. with **`<p>`** and **`</p>`**, you can type the tag in here. Case sensitive, too.

In many cases you may consider that defining a paragraph end will suffice (considering everything up to it to be part of the preceding one). Much HTML text does not consistently distinguish between paragraph starts and ends.

Note that spoken texts in the BNC use **`</u>`** instead of **`</p>`**, but you can leave **`</p>`** here as WordSmith will use **`</u>`** instead if the text has no **`</p>`** in it.

See also: Tagged Text 131, Stop Lists 120, Choosing a new language 7. Processing text in Chinese etc.

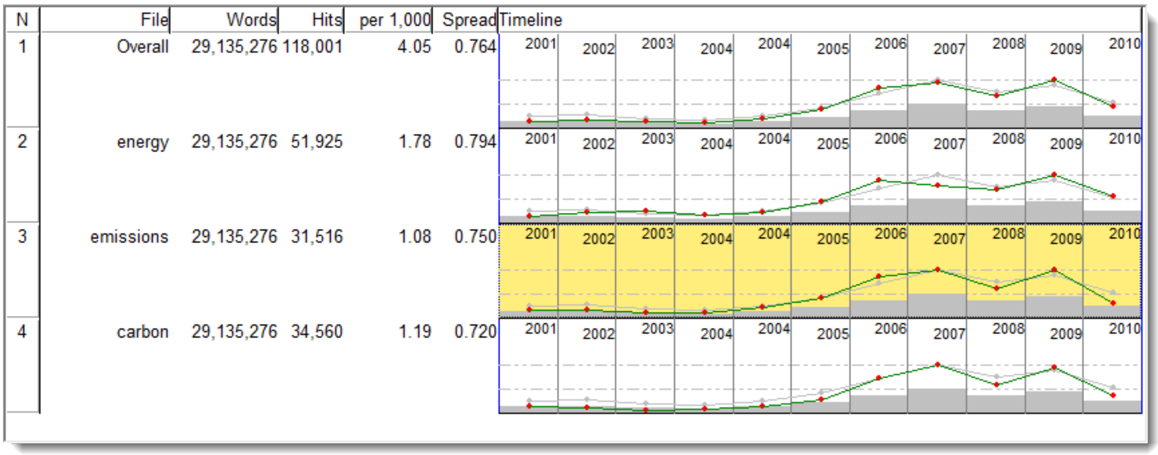
5.37 text dates and time-lines

The point of it ...

The idea is to be able to treat your text files *diachronically* -- that is, studying change through time.

You might want a concordance, for example, to be ordered by the text date. Or you might be interested in knowing when a certain word first appeared in your corpus and whether it gained popularity in succeeding years. Or whether the collocates of a word like **`web`** changed before 1990 and after.

This screenshot shows a time-line based on concordancing **`energy/emissions/carbon`** in about 30 million words of UK newspaper text dealing with climate change, 2000-2010.



N	File	Words	Hits	per 1,000	Spread	Timeline
1	Overall	29,135,276	118,001	4.05	0.764	
2	energy	29,135,276	51,925	1.78	0.794	
3	emissions	29,135,276	31,516	1.08	0.750	
4	carbon	29,135,276	34,560	1.19	0.720	

The first line shows overall data where all results on three search-terms are merged.

Concordance hits are represented as a graph with green lines and little red blobs for each time period.

The grey rectangles and the grey graph line both represent the same background information, namely the amount of word-data searched. The difference is merely that the grey line is twice as high as the rectangles below it.

The number of hits in each year is mostly roughly proportional to the amount of text being examined, though in 2006 and 2009 for the term **`emissions`** it seems that the hit rate was slightly higher. In the first half of the decade **`carbon`** was rather under-mentioned in proportion to the amount of climate-change data being studied.

See also choosing text files: setting file dates 48

5.38 window management

The main WordSmith Tools Controller 4 will be at the top left corner of your screen, half the screen width and half the screen height in size. Other Tools will appear in the middle. Each Tool main window will come just below any previous ones.

Make use of the Taskbar (or Alt-tab, which helps you to switch easily from one window to the next).

"Start another Concord window"?

You will see this if you already have a window of data and press *New* to start another concordance. You can have any number of windows open for each Tool, each with different data.

minimising, moving and resizing windows

All windows can be stretched or shrunk by putting the mouse cursor at one edge and pulling. They can be moved most easily by grabbing the top bar, where the caption is, and pulling, using the mouse. You can minimise a window: it becomes an icon which you restore by clicking on it. If you maximise it, it will fill the entire screen of the Tool concerned. These are standard Windows functions. It's okay to minimise the main Controller 4 window when using individual Tools.

tile and cascade

You can *Tile* or *Cascade the Tools* from the main **WordSmith Tools** program.

restore last file

A convenience feature: the last file you saved or retrieved will by default be restored when you re-enter WordSmith Tools. I've kept it to one only to avoid screen clutter! This feature can be turned off temporarily via a settings option or permanently in **`wordsmith6.ini`** (in your **`Documents\wsmith6`** folder). You can also generally access your last saved result in any Tool by right-clicking

and choosing last file:

5.39 word clouds

The point of it...

Many of the lists in WordSmith offer a word cloud feature similar to those you have probably seen on the web. The idea is to promote pattern-noticing 450.

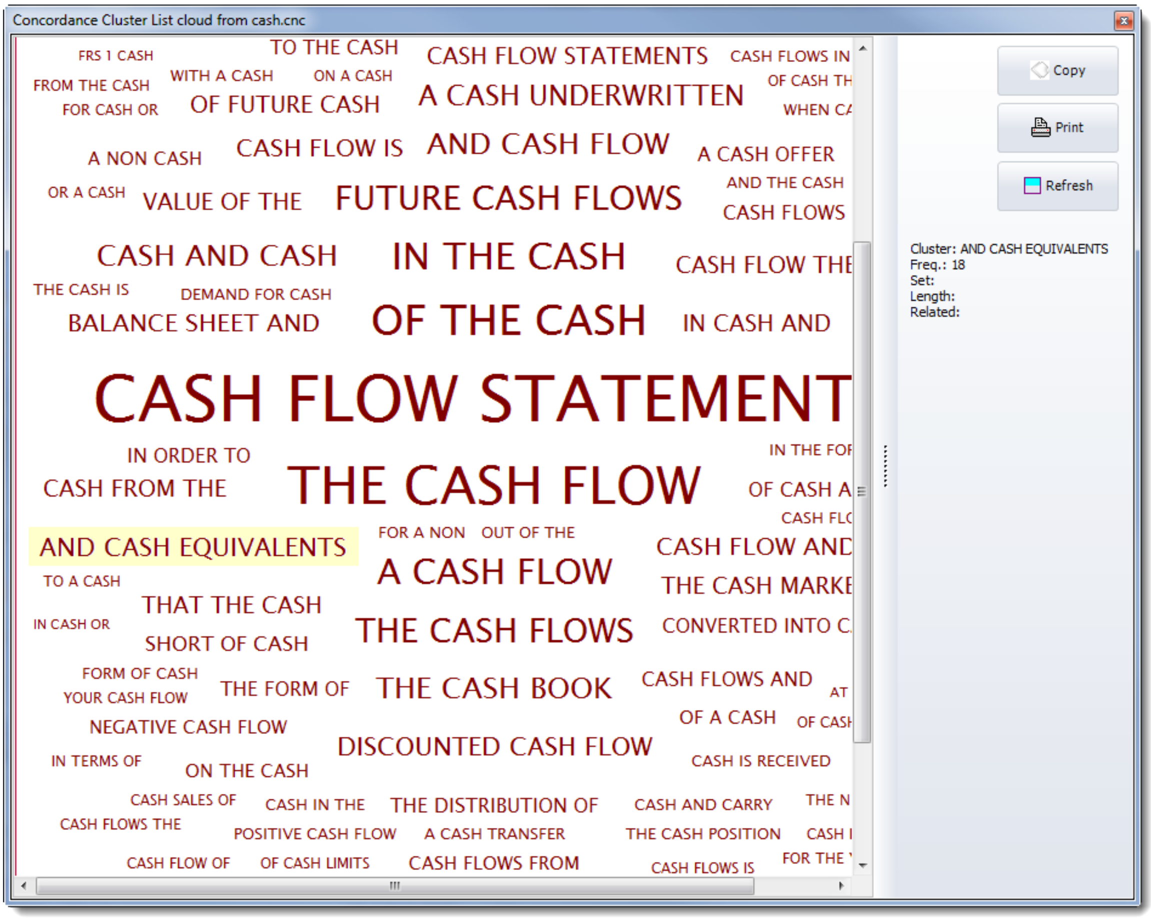
How to get here

This function is accessed from the Compute menu, sub menu-item Word Cloud () in the various Tools.

Examples

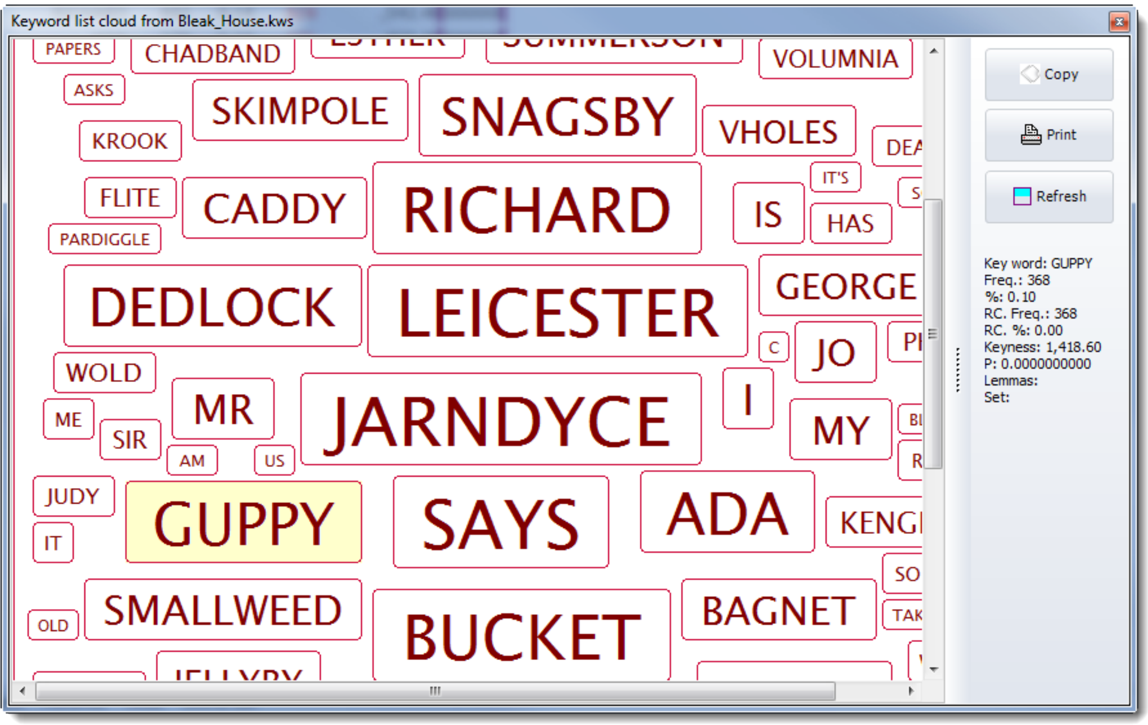
In this case of collocates of 189 ***cash*** 189 you can get a word cloud based on any column of data.

With Concord clusters based on **cash**, this example was computed:



Cluster: AND CASH EQUIVALENTS
Freq.: 18
Set:
Length:
Related:

In the case of key words, you can get something like this:



Key word: GUPPY
Freq.: 368
%: 0.10
RC. Freq.: 368
RC. %: 0.00
Keyness: 1,418.60
P: 0.0000000000
Lemmas:
Set:

In this case the word cloud was based on the key words of a novel, *Bleak House* (Charles Dickens). The highlighted word **Guppy** is the name of one of the characters and details of this word are shown to the right.

What you can see and do

The *Copy* and *Print* buttons do what their names suggest. The *Refresh* button recalculates the cloud, e.g. after you have deleted items in the original data.

As your mouse hovers over a word in the cloud you get details of that individual word.

You can change the word cloud settings in the main Colours setting in the Controller.

The font sizes range from a minimum of 8 to a maximum of 40 depending on the range of values in your data. The font 78 is the one you may choose for any of your standard displays.

5.40 zap unwanted lines

To restore the correct order to your data after editing it a lot or marking lines for deletion, press the Zap button (⚡ or Ctrl+Z). This will permanently cut out all lines of data which you have deleted (by pressing Del) unless you've restored them from deletion (Ins).

In the case of a word list, it will also re-order the whole file in correct frequency order. Any deleted entries are lost at this stage. Any which have been assigned as lemmas of head words may still be viewed, before or after saving. However, after zapping, lemmas can no longer be undone.

In the case of a concordance, you may wish the list of filenames to be re-computed to reflect only the files still referred to in your concordance. To do that, choose *Compute* | *Filenames*.

See also : reduce data to N entries 70.

Tags and Markup

Section VI

6 Tags and Markup

6.1 overview

What is markup for?

Marked up text is text which has extra information built into it with tags, e.g. "We<pronoun> like<verb> spaghetti<noun>.<end of sentence>". You may wish to concordance words or tags…

You may wish to see this additional information or ignore it, so that you just see the plain text ("We like spaghetti."). WordSmith has been designed so that you can choose what to ignore and what to see.

You may want to translate HTML or SGML 435 tags or entity references: if your text has **`&Eacute;`** you probably want to see **É**.

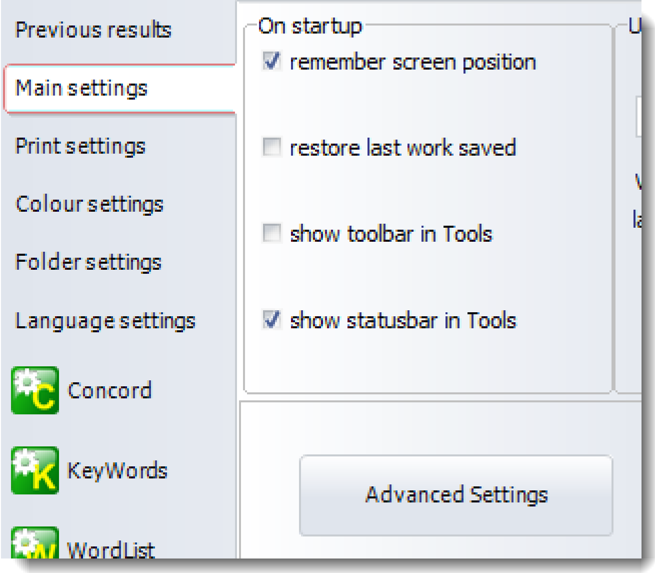
You may wish to select within text files, e.g. cutting out a header or getting only the conclusions, instead of using the whole text.

And you might want to get WordSmith to choose only files meeting certain criteria, e.g. having "**`sex=f`**" in a text file header section, where the speaker is a woman.

You can see the effect of choosing tags if you select the Choose Texts option, then press the View 390 button. Any retained tags will be visible, and ignored tags replaced by spaces.

Tags and Markup Settings

… are accessed via an Advanced Settings button in the Controller



☑ remember screen position

☐ restore last work saved

☐ show toolbar in Tools

☑ show statusbar in Tools

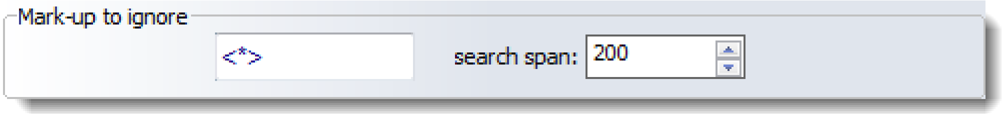
See also: Guide to handling the BNC, Handling Tags 132, Making a Tag File 141, Showing Nearest Tags in Concord 199, Concord Sound and Video 212, Tag Concordancing 198, Types of Tag 145, Viewing the Tags 390, Using Tags as Text Selectors 134, Tags in WordList 311, XML text 153

6.2 choices in handling tags

ignore all tags

Specify all the opening and closing symbols in *Main Settings* | *Advanced* | *Tags* |*Mark-up to ignore* and such tags will be simply left out of word lists and concordances, as if they weren't in the original text files.

example :



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

<*> This will cut out all wording starting at each **<** symbol and ending at the next **>** symbol (up to 200 characters apart). (You can put more than one pair of brackets, e.g. <*>;[*] if you like.)

ignore some tags and retain others

If you want to ignore some but retain others, you will need to prepare a tag file 141 which lists all those you want to keep. These will then appear in your word lists and concordances.

You get WordSmith Tools to read this text file in by choosing the Tag File menu option under Settings. Such tags will then be incorporated into your word lists, concordances, etc. as if they were ordinary words or suffixes.

example: supposing you've set **`<*>`** as "tags to ignore", but listed **`<title>`**, **`<body>`** and **`<conclusion>`** as tags to retain in your tag file, WordSmith will keep any instances of **`<title>`**, **`<body>`** or **`<conclusion>`** in your data but will ignore **`<introduction>, <Ulan Bator>, <threat>`**, etc.

Tags to retain will only be active if there's a file name visible and you have pressed the *Load* or *Clear* button. If you press *Load*, you will see which tags have been read in from the tag file.

translate entity references into other characters

If you use XML, SGML or HTML 435 tagged text, you may want to translate symbols. For example, SGML, XML, HTML use **`&mdash;`** instead of a long dash. To do this, first prepare a Tag File 141 which contains the strings you want to translate. Then choose *Main Settings* | *Advanced* | *Tags & Markup* | *Entity File (entities to be translated)* and choose your entity file. WordSmith will then translate any entity references in this file into the corresponding characters.

to load up these tag files automatically

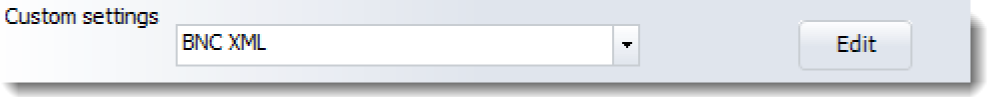
See Custom Settings 133.

See also: Guide to handling the BNC, Overview of Tags 131, Making a Tag File 141, Showing Nearest Tags in Concord 199, Tag Concordancing 198, Types of Tag 145, Viewing the Tags 390, Using Tags as Text Selectors 134, Tags in WordList 315

6.3 custom settings

Custom Tagsets

In the main *Settings* | *Tags & Markup* window, you may see custom settings choices like this.



Custom settings: BNC XML

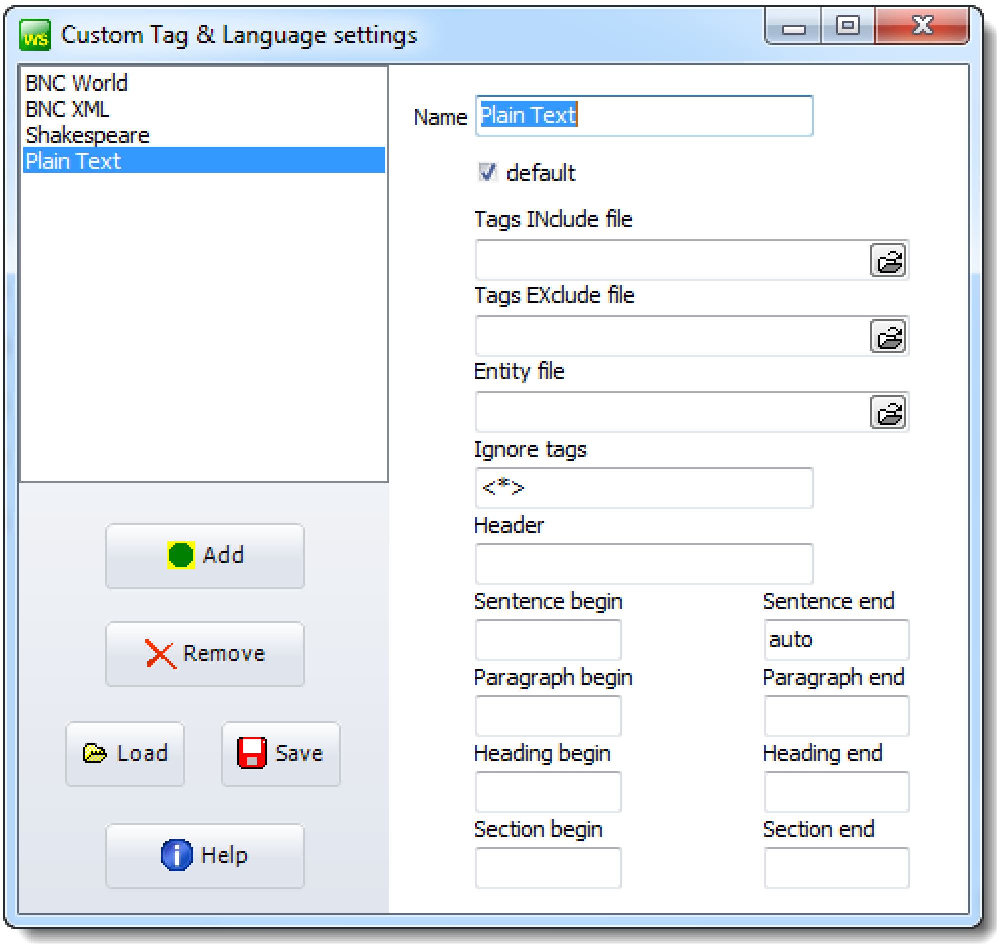
The point of it...

The point of this choice is to change a whole series of settings according to the type of corpus you wish to process.

When you change the setting above, any valid data as explained below will get loaded into your defaults.

How to do it

Press the Edit button to create or edit the custom settings (the file is called **`custom_tag_settings.xml`** and it'll get saved in your **`Documents\wsmith6`** folder).



Custom Tag & Language settings

BNC World
BNC XML
Shakespeare
Plain Text

Name: Plain Text

☑ default

Tags INclude file:

Tags EXclude file:

Entity file:

Ignore tags: <*>

Header:

Sentence begin:

Sentence end: auto

Paragraph begin:

Paragraph end:

Heading begin:

Heading end:

Section begin:

Section end:

To start a new set, press *Add* and give a suitable name (such as **`Shakespeare`** for processing the Shakespeare corpus).

Fill in the boxes and press Save.

- All boxes will have leading and trailing spaces removed.
- Use **`auto`** for automatic processing e.g. of sentence ends.
- Checking the *default* box means that this set gets chosen by default and any tag or entity files will get automatically loaded for you.

See also : Tags as text selectors 134

6.4 tags as selectors

Defaults

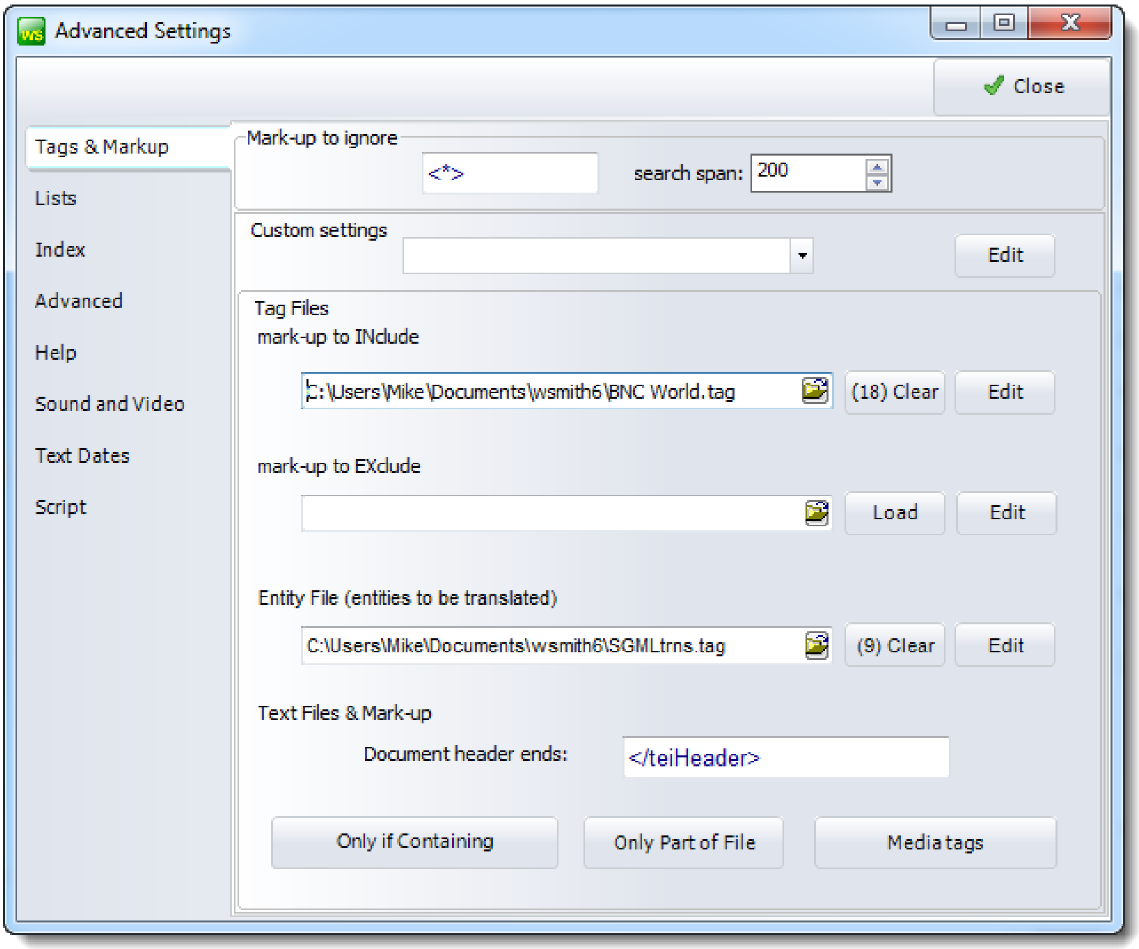
The defaults are: select *all* sections of *all* texts selected in Choose Texts 44 but cut out all angle-bracketed tags.

Custom settings

There are various alternatives in this box which help your choices with the boxes below. Choosing *British National Corpus World Edition* (as in the screenshot) will for example automatically put **`</teiHeader>`** into the Document header ends box below. You can also edit the options 133 and their effects.

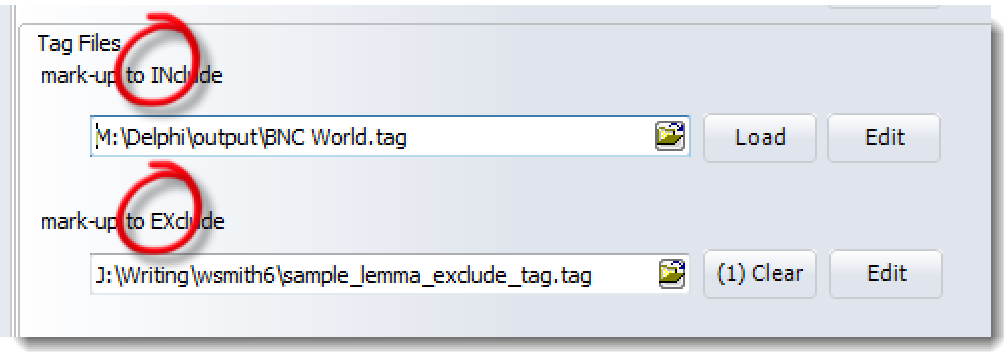
Markup to ignore

If you want to cut out unwanted tags eg. in HTML 435 files, leave something like **`< >`** or **`[ ]`** or **`< >;[ ]`** in *Markup to ignore*. The "search-span" means how far should WordSmith look for a closing symbol such as **`>`** after it finds a starting symbol such as **`<`**. (The reason is that these symbols might also be used in mathematics.)



Mark-up to ignore: <*>
search span: 200
mark-up to INclude: C:\Users\Mike\Documents\wsmith6\BNC World.tag
Entity File (entities to be translated): C:\Users\Mike\Documents\wsmith6\SGMLtrns.tag
Document header ends: </teiHeader>

Markup to INclude or EXclude



Tag Files

mark-up to INclude

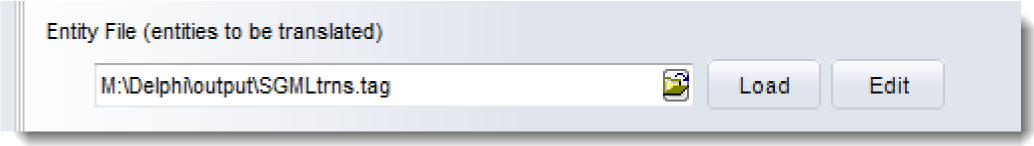
M:\Delphi\output\BNC World.tag

mark-up to EXclude

J:\Writing\wsmith6\sample_lemma_exclude_tag.tag

See Making a Tag File 141.

Entity file



Entity File (entities to be translated)

M:\Delphi\output\SGMLtrns.tag

See Making a Tag File 141.

Text Files and Mark-up

However, you can get **WordSmith** to use tags to select one section of a text and ignore the rest. This is "selecting within texts". You can also select *between* texts: that is, get **WordSmith** to look within the start of each text to see whether it meets certain criteria.

These functions are available from *Main Settings* | *Advanced* | *Tags* | *Only If Containing* 137 *or Only Part of File* 138.

Document Header

When you process a set of texts usually containing a standard header (e.g. a copyright notice) you may wish to remove it automatically.

Ensure that some suitable tag is specified as above in the **`</teiHeader>`** example. (If you choose Custom Settings above, you will get suitable choices automatically.) The process cuts by looking for the *Document header ends* mark-up and deleting all text to that point. (If you have a header repeated in the same text file, WordSmith will need to be told what mark-up is used for *Document header starts* too, and you will need to choose Only Part of File 138 to get such headers removed.)

For more complex searches, you might want to choose the Only If Containing 137 or Only Part of File 138 buttons visible above.

The order in which these choices are handled

If you choose either to select either between or within texts, WordSmith will check that each text file meets your requirements, before doing your concordance, word list, etc. It will

1. Select between files 137 to check whether it contains the words you've specified;
2. Cut out any section specified as a "section to cut 138";
3. If there are "sections to keep 138", cut out everything which is not within them;

4. Cut start of each line 138, if applicable;
5. Process any entity references you want to translate 132;
6. Ignore 132 any tags not to be retained (see the "Mark-up to ignore" section of the screenshot above).

See also: Overview of Tags 131, Making a Tag File 141, Tag Handling 132, Tag Concordancing 198, Showing Nearest Tags in Concord 199, Viewing the Tags 390, Types of Tag 145, Guide to handling the BNC, XML text 153

6.5 only if containing...

The point of it

You might want to process only the speech of elderly men, or only advertising material, or only classroom dialogues. This function allows WordSmith to search through each text, e.g. in text headers, ensuring that you get the right text files and skip any irrelevant ones.

Suppose you have a large collection of texts (e.g. the British National Corpus) and you cannot remember which of them contain English spoken by elderly men.

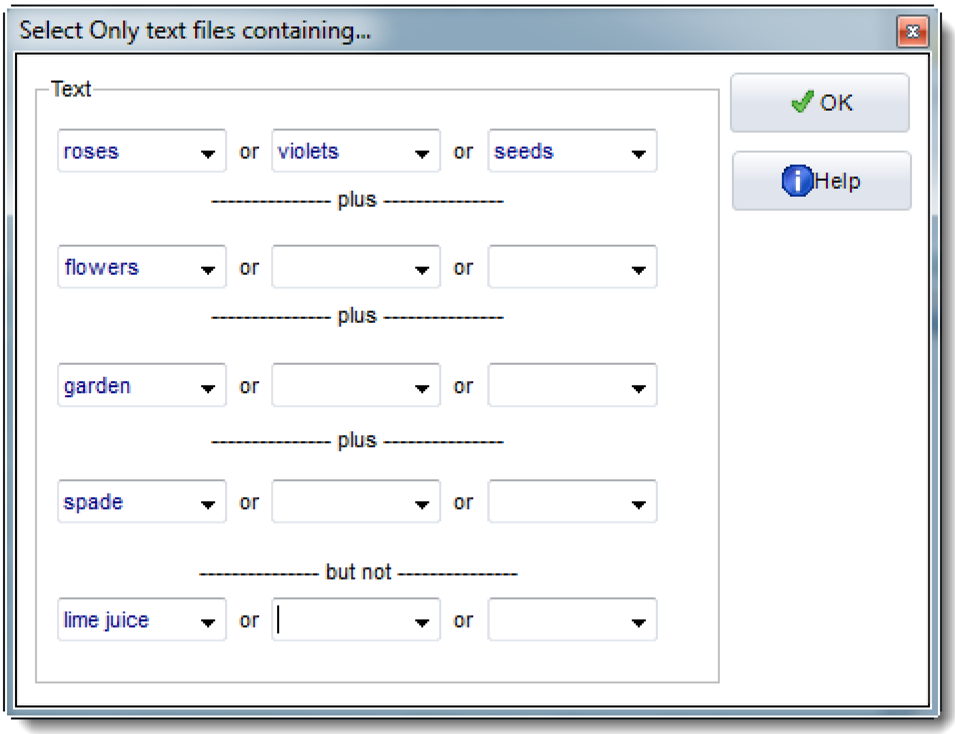
Knowing that the BNC uses **`stext>`** for spoken texts, **`sex=m`** for males, **`age=5`** for speakers aged 60 or more, you can get WordSmith to filter your text selection. It will search through the whole of every text file (not just the tags or header sections, in fact the first 2 megabytes 437 of the file) to check that it meets your requirements.

You can specify up to 15 tags, each up to 80 characters in length. They will be case sensitive (i.e. you will get nothing if you type **`Age=5`** by mistake).

Horizontally, the options represent combinations linked by "or". Vertically, the combinations are "and" links. The bottom set represents "but not" combinations.

After your text files have been processed, you will be able to see which met your requirements in the Text File choose window 44 and can save the list for later use as favourites 50.

Examples:



roses or violets or seeds

plus

flowers or or

plus

garden or or

plus

spade or or

but not

lime juice or or

You only want text files which contain either **roses** or **violets** or **seeds**, and **flowers** must be present too, so must **garden** and **spade** But you do not want **lime juice** to be present in the text.

If you want **book** or **hotel** but only if they're not in a text file containing **publish** or **Booker Prize**: write **book** into the first box, **hotel** in the box beside it, and **publish*** and **Booker *** in the first two boxes in the bottom row.

See also: Tags as Selectors 134, Selecting within texts 138, Extracting text sections 359, Filtering your text files 360 using Text Converter 354, Guide to handling the BNC

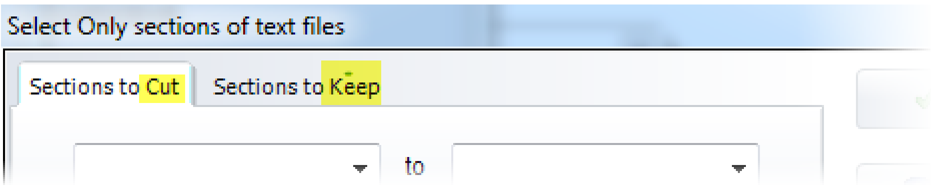
6.6 part of file:selecting within texts

The point of it

The aim is to let you get WordSmith to process only specific parts of your text files, getting rid of chunks you're not interested in.

Cut out or Keep?

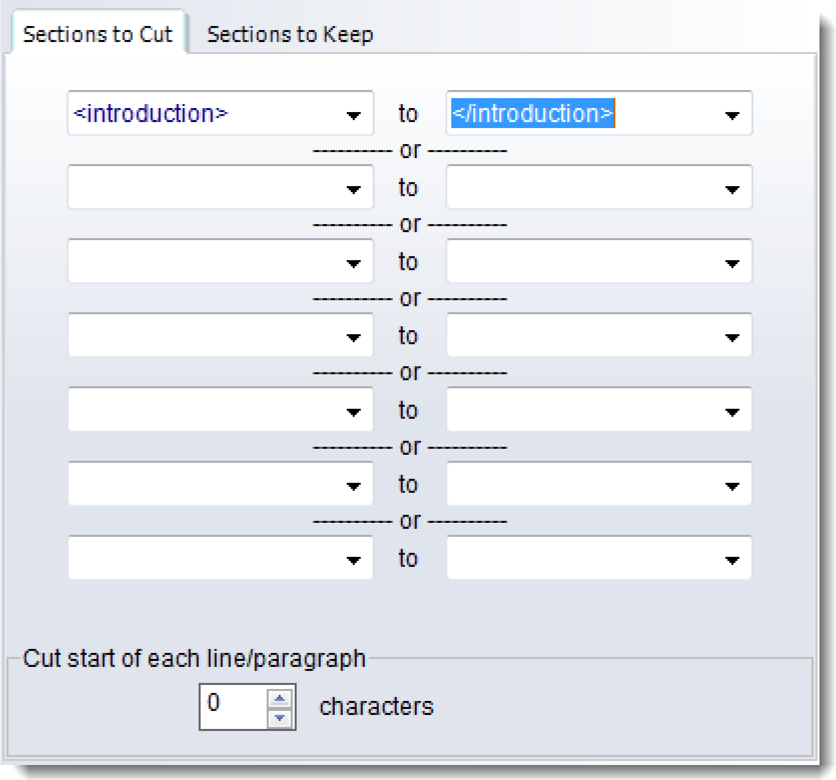
Press the *Cut* or *Keep* tab to choose to cut out certain sections, and/or only to use certain sections.



Sections to Cut

Note: if you only want to remove a document header such as **`</header>`**, it is easier to do that in the general tag settings 134, section Document Header.

For more complex choices, you may here specify what is to be cut, where it starts (for example **`<introduction>`**) and where you want to cut to (e.g. **`</introduction>`**). You can choose to cut out up to 7 different and separate sections (**`<HEAD>`** to **`</HEAD>`** or **`<BODY>`** to **`</BODY>`**). This function is case-sensitive and cuts out any section located as many times as it is found within the whole text.



Sections to Cut — Sections to Keep

<introduction> to </introduction>

Cut start of each line/paragraph: 0 characters

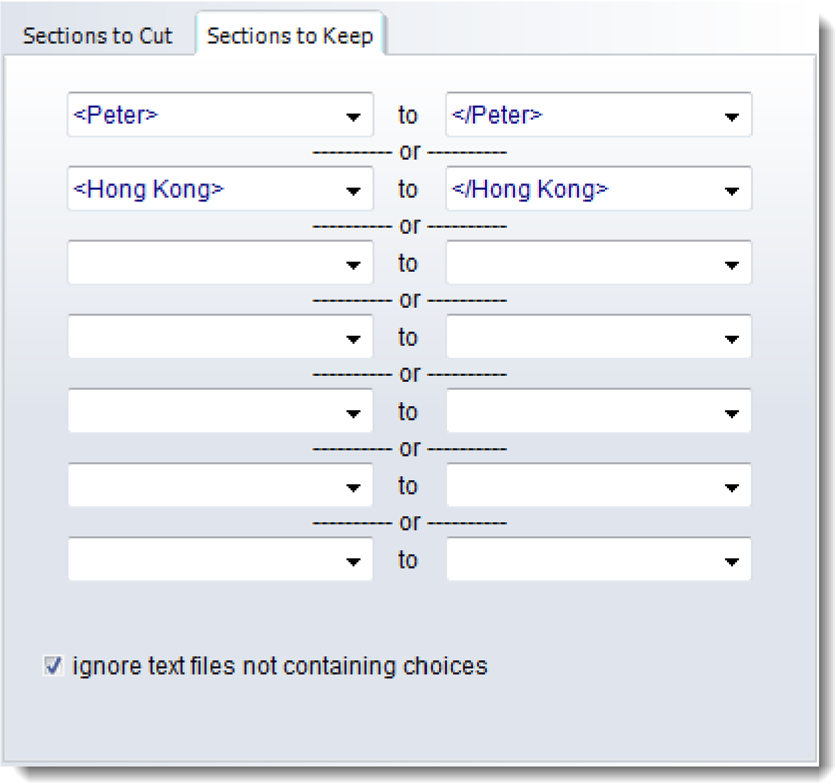
Cut start of each line/paragraph

The point of this is that some corpora (e.g. LOB) have a fixed number of line-detail codings at the start of each line. Here you want to cut them out (that is, after every <Enter>). Choose the number of characters to cut, up to 100; the default is 0. Use **`-1`** if you want to cut everything up to the first alphabetical character at the start of each line, and **`-2`** to cut everything up to the first tab.

Sections to Keep (contexts)

You want to select just one or two sections of the text and cut out the rest. Specify one tag to define the desired start, and one to specify the end, e.g. **`<Intro>`** to <**`Body>`**

(these would analyse only text introductions), or <**`Mary>`** to **`</Mary>`** (these would get all of Mary's contributions in the discourse but nothing else).



<Peter> to </Peter>

or

<Hong Kong> to </Hong Kong>

☑ ignore text files not containing choices

Here we have chosen to use 2 different sections, **`<Peter>`** to **`</Peter>`** to get the sections spoken by Peter and **`<Hong Kong>`** to **`</Hong Kong>`** to get the sections marked up as referring to Hong Kong as well.

Naturally you must be sure that there is something unique like a **<** or **>** symbol to define each section. This function is case sensitive (so it would not find **`<PETER>`**).

If you used **`<H1>`** to **`</H1>`** with this function in HTML 435 text you'd get all the major headings in your texts, however many, but nothing else.

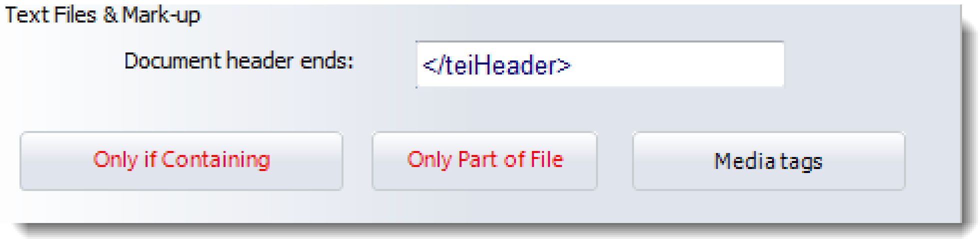
The "off" switch doesn't have to look like the "on" switch -- you could keep, for example, **`<INTRO>`** to **`</BODY>`** and thereby cut out the conclusion if that comes after the **`</BODY>`**.

Ignore text files not containing choices

If this is checked, your text files will be examined to ensure they contain the mark-up for sections to keep (here **`<Peter>`** and **`<Hong Kong>)`**.

OK

Once you've pressed OK, you will see that WordSmith knows you want only certain parts of each file because the *Only Part of File* button goes red (as will the *Only if Containing* button if there were sections to keep and the *Ignore text files not containing choices* box was checked).



Text Files & Mark-up

Document header ends: </teiHeader>

See also: Tags as Selectors 134, Only if containing <x> 137, Guide to handling the BNC.

6.7 making a tag file

Tag Syntax

Each tag is case sensitive.

Tags conventionally begin with **<** and end with **>** but the first & last characters of the tag can be any symbol.

You can use

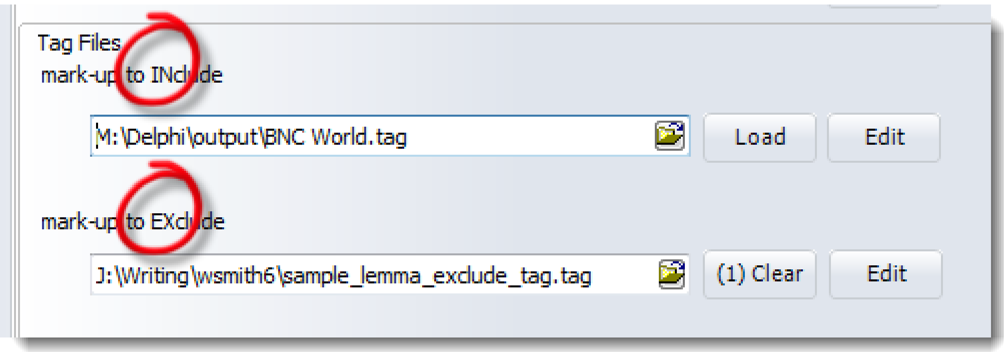
***** to mean any sequence of characters;
? to mean any one character;
# to mean any numerical digit.

Don't use **[** to insert comments in a tag file, since **[** is useful as a potential tag symbol. You can use # to represent a number (e.g. **<h#>** will pick up **<h5>, <h1>**, etc.). And use **?** to represent any single character (**<?>** will pick up **<s>, <p>**, etc.), or * to represent any number of characters (e.g. **<u*>** will pick up **<u who=Fred>, <u who=Mariana>**, etc.). Otherwise, prepare your tag list file in the same way as for Stop Lists 120.

Use **notepad** or any other plain text editor, to create a new **.tag** file. **Write one entry on each line**.

Any number of pre-defined tags can be stored. But the more you use, the more work WordSmith has to do, of course and it will take time & memory ...

Mark-up to EXclude

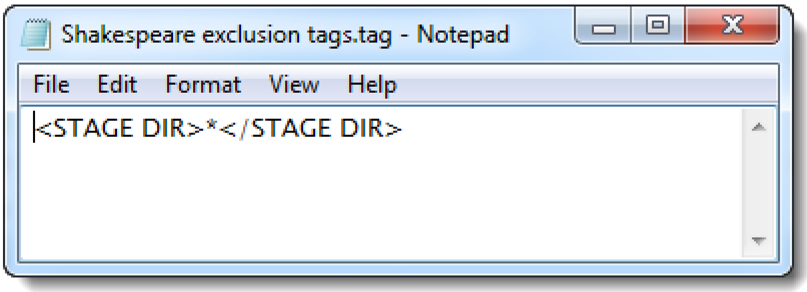


Tag Files
mark-up to INclude
M:\Delphi\output\BNC World.tag
mark-up to EXclude
J:\Writing\wsmith6\sample_lemma_exclude_tag.tag

A tag file for stretches of mark-up like this **`<SCENE>A public library in London. A bald-headed man is sitting reading the News of the World.</SCENE>`**

where you want to exclude the whole stretch above from your concordance or word list, e.g. because you're processing a play and want only the actors' words. Mark-up to exclude will cut out the whole string from the opening to the closing tag inclusive.

For the Shakespeare corpus, a set of tags to EXclude might be used.



Shakespeare exclusion tags.tag - Notepad

```
<STAGE DIR>*</STAGE DIR>
```

(The idea is not to process any stage directions when processing the Shakespeare corpus.)

The syntax requires **`></`** or **`>*</`** to be present.

Legal syntax examples would be:

```
<SCENE></SCENE>
<SCENE>*</SCENE>
<SCENE #>*</SCENE>
<HELLO?? #>*</GOODBYE>
```

(In this last example it'll cut only if **`<HELLO`** is followed by 2 characters, a space and a number then **`>`**, and if **`</GOODBYE>`** is found beyond that.)

```
<SCENE>*
</SCENE>
```

won't work, because both parts of the tag must be on the same line.

```
<SCENE>*<\SCENE>
```

won't work, because the slash must be **`/`**.

With your installation you will find (**`Documents\wsmith6\sample_lemma_exclude_tag.tag`**) included, which cuts out lemmas if constructed on the pattern **`<lemma tag="*>*</lemma>`**, i.e. with the word **`tag`**, an equals sign and a double-quote symbol, regardless of what is in the double-quotes.

Mark-up to INclude

A tag file for tags to retain contains a simple list of all the tags you want to retain. Sample tag list files for BNC handling (e.g **`bnc world.tag`**) are included with your installation (in your **`Documents\wsmith6`** folder): you could make a new tag file by reading one of them in, altering it, and saving it under a new name.

Tags will by default be displayed in a standard tag colour 60 (default=grey) but you can specify the foreground & background for tags which you want to be displayed differently by putting

/colour="foreground on background"
e.g. **`<noun> /colour="yellow on red"`**
Available colours:
'Black','White','Cream',
'Red','Maroon',
'Yellow',
'Navy','Blue','Light Blue','Sky Blue',
'Green','Olive','Dollar Green','Grey-Green','Lime',
'Purple','Light Purple',
'Grey','Silver','Light Grey','Dark Grey','Medium Grey'.

The colour names are not case sensitive (though the tags are). Note UK spelling of "grey" and "colour".

Also, you can put "/play media" if you wish a given tag, when found in your text files, to be able to attempt to play a sound or video file 147. For example, with a tag like

```
<sound *> /colour="blue on yellow" /play media
```

and a text occurrence like

```
<sound c:\windows\Beethoven's 5th Symphony.wav>
```

or

```
<sound http://www.political_speeches.com/Mao_Ze_Dung.mp3>
```

you will be able to choose to hear the .wav or .mp3 file 212.

Finally, you can put in a descriptive label, using /description "label" like this:

```
<w NN*> /description "noun" /colour="Cream on Purple"
<ABSTRACT> /description "section"
<INTRODUCTION> /description "section"
<SECTION 1> /description "section"
```

Tagstring_only tags

You can also define two tags as ones you want to use to mark the beginnings and ends of what will be shown in a concordance using **`/tagstring_only`** as a signal. For example, if concordancing text containing titles marked out with **`<title>`** and **`</title>`**, you may want to see only the title text. You'd include in the tag file

```
<title> /tagstring_only
</title> /tagstring_only
```

To get Concord to show only the text between these two, choose *View* | *Tag string only* in Concord's menu.

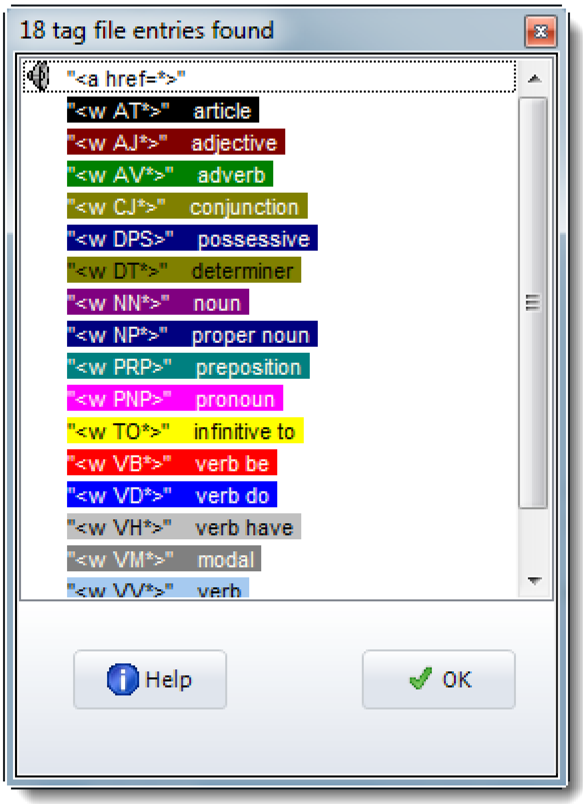
Section tag

In the examples using "section", Concord's "Nearest Tag" will find the section however remote in the text file it may be.

This is particularly useful e.g. if you want to identify the speech of all characters in a play, and have a list of the characters, and they are marked up appropriately in the text file.

```
<Romeo> /description "section"
<Mercutio> /description "section"
<Benvolio> /description "section"
```

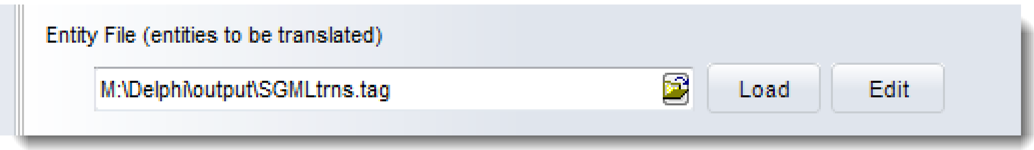
Here is an example of what you see after selecting a tag file and pressing "Load". The first tag is a "play media" tag, as is shown by the icon. You can see the cream on purple colour for nouns too. The tag file (**BNC World.tag**) is included in your installation.



18 tag file entries found

"<a href=*>"
"<w AT*>" article
"<w AJ*>" adjective
"<w AV*>" adverb
"<w CJ*>" conjunction
"<w DPS>" possessive
"<w DT*>" determiner
"<w NN*>" noun
"<w NP*>" proper noun
"<w PRP>" preposition
"<w PNP>" pronoun
"<w TO*>" infinitive to
"<w VB*>" verb be
"<w VD*>" verb do
"<w VH*>" verb have
"<w VM*>" modal
"<w VV*>" verb

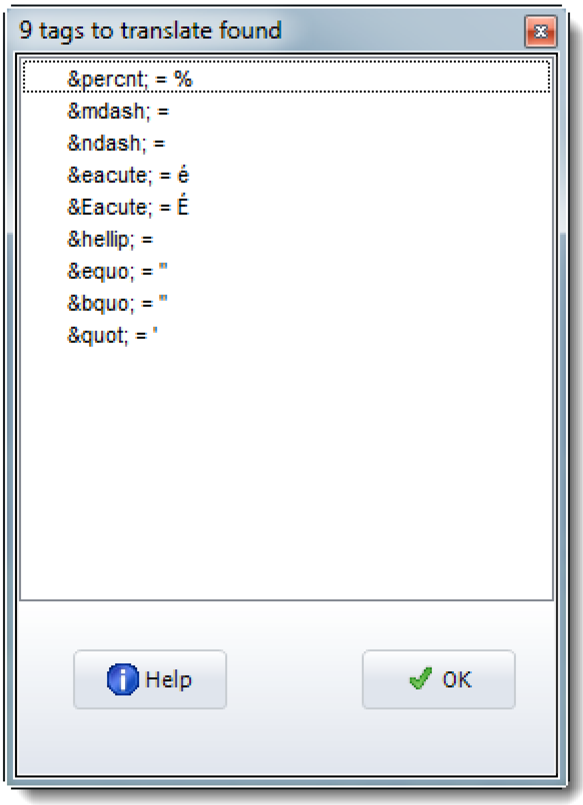
Entity File (entities to be translated)



Entity File (entities to be translated)

M:\Delphi\output\SGMLtrns.tag

If you load it you might see something like this:



9 tags to translate found

&percnt; = %
&mdash; =
&ndash; =
&eacute; = é
&Eacute; = É
&hellip; =
&equo; = "
&bquo; = "
&quot; = '

A tag file for translation of one entity reference into another uses the following syntax: entity reference to be found + space + replacement. Examples:

```
&Eacute; É
&eacute; é
```

In the screenshot above, the sample tag file for translation (**`Documents\wsmith6\sgmltrns.tag`**) which is included with your installation has been loaded. You could make a new one by reading it in, altering it, and saving it under a new name.

See also: Overview of Tags 131, Handling Tags 132, Showing Nearest Tags in Concord 199, Tag Concordancing 198, Types of Tag 145, Viewing the Tags 390, Using Tags as Text Selectors 134, Guide to handling the BNC.

6.8 tag-types

You will need to specify how each tag type starts and ends, and you should be consistent in usage. Restrict yourself to symbols which otherwise do not appear in your texts.

eight special markers

Eight kinds of marker may be marked as significant for word lists: those which represent starts and ends of headings 147, sections 147, sentences 147 and paragraphs 147. Type these in the appropriate spaces when selecting Text Characteristics 124.

tags within 2 separators 427

These tags are often used to signal the part of speech of each word; they're also widely used in HTML, XML, SGML 435 for "switches", e.g. **`<H1>`** to switch on Heading 1 style and **`</H1>`** to switch it off again. You should use the same opening and closing symbols, usually some kind of brackets, for all your tags (as the British National Corpus does using SGML or XML 435 markup): **`<Noun>,<Verb>,<Pronoun>`**.

entity references

HTML, XML and SGML 435 use so-called entity references for symbols which are outside the standard alphabet, e.g. **`&eacute;t&eacute`** which represents **`été`**.

Specify these two types of markup by choosing Settings/Tag Lists, or Settings/Text Characteristics/ Tags. You will then see a dialogue box offering Text to Ignore and a Browse button.

The Tags to Ignore 132 option allows you to specify tags which you do not want to see in the concordance or word list results.

The Tags to be INcluded 141 option allows you to specify a tag file, containing tags which you do want to see in the concordance or word list results.

The Tags to be EXcluded 141 option allows you to specify a different tag file, containing stretches of tags which you want to find and remove in the concordance or word list results.

The Tags to be Translated 132 option allows you to specify entity references which you want to convert on the fly, such as **`&eacute`**.

multimedia markers

Text files can be tagged for reference to sound or video files which you can hear or see. For example, a text might contain something like this: **`blah blah blah ...<a href=http://gandalf.hit.uib.no/c/l/32401-1.mp3> blah blah`** etc. A concordance on **`blah blah`** could pick up the tag so you can hear the source mp3 file. See defining multimedia tags 147.

See also: Overview of Tags 131, Handling Tags 132, Making a Tag File 141, Showing Nearest Tags in Concord 199, Tag Concordancing 198, Viewing the Tags 390, Using Tags as Text Selectors 134, Concord Sound and Video 212, Guide to handling the BNC.

(A particular sub-variety of tags within 2 separators sometimes used is tags with underscores at the left and space at the right like this

`He_PRONOUN entered_VERB the_DET room_NOUN`.

To process these, you will need to declare the underscore a valid character 125, or else convert your corpus 368 to a format like.

`<PRONOUN>He <VERB>entered <DET>the <NOUN>room`.)

6.9 start and end of text segments

WordSmith attempts to recognise 4 types of text segment: sentences, paragraphs, headings, sections. Processing is case sensitive. You can use **`<Enter>`** and **`<Tab>`** as strings representing an end of paragraph or a tab in your texts. For sentence ends, **`auto`** is another option.

Define these in your language settings 81.

Sentences

For example, **`<s>`** might represent the beginning of a sentence and **`</s>`** the end. If you leave the choice as **`auto`**, ends of sentences are determined by according to the definition of a sentence 426 which gives a approximation. (There is no 100% accurate way 390 of handling sentence recognition.)

Paragraphs

For example, **`<p *>`** or **`<p>`** might represent the beginning of a paragraph and **`</p>`** the end.

Headings

For example, **`<head>`** might represent the beginning and **`</head>`** the end. Note that the British National Corpus marks sentences within headings. Eg.

```
<head>
<s n="2"><w NN1>Introduction
</head>
```

in text **`HXL`**. It seems odd for the one word **`Introduction`** to count as a sentence, so WordSmith does not use sentence-tags within headings.

Sections

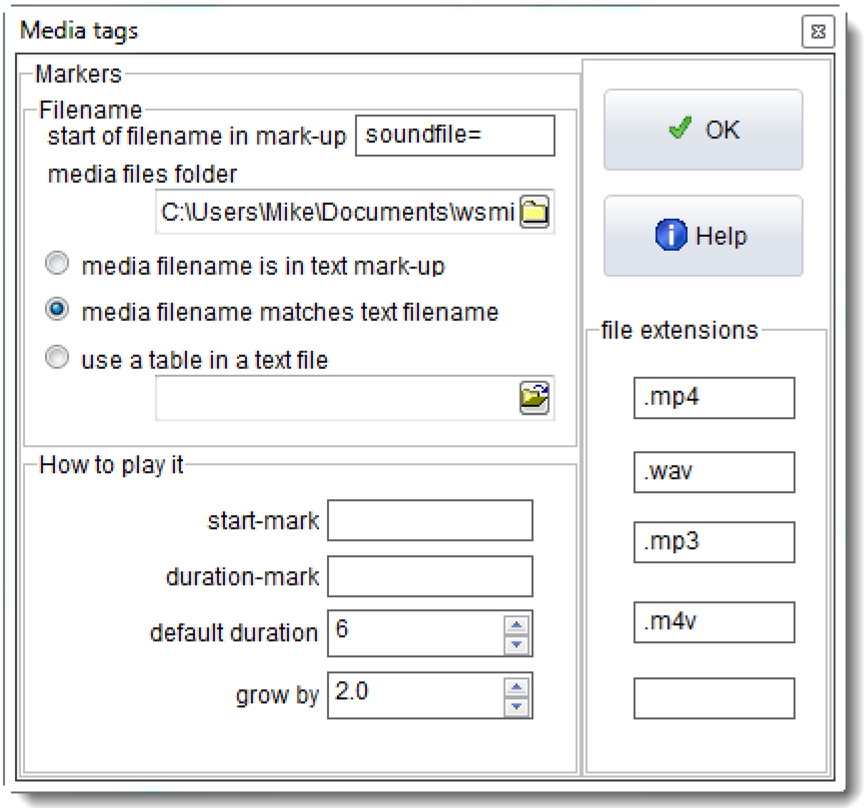
For example, **`<section *>`** might represent the beginning and **`</section>`** the end.

Each of these is counted preferably when its closing tag such as **`</s>, </p>`** etc. is encountered. If there are no closing **`</p>`** tags in the entire text then paragraphs will be counted each time the opening paragraph tag is found.

See also: Overview of Tags 131, Handling Tags 132, Showing Nearest Tags in Concord 199, Tag Concordancing 198, Types of Tag 145, Viewing the Tags 390, Using Tags as Text Selectors 134, Guide to handling the BNC.

6.10 multimedia tags

In this screenshot you see an example of how to define your multimedia tags. This is accessed from *Main Settings* | *Advanced* | *Tags* | *Media Tags*.



Media tags

Markers

Filename

start of filename in mark-up: soundfile=

media files folder: C:\Users\Mike\Documents\wsmi

○ media filename is in text mark-up

● media filename matches text filename

○ use a table in a text file

How to play it

start-mark

duration-mark

default duration: 6

grow by: 2.0

file extensions: .mp4, .wav, .mp3, .m4v

File Extensions

The file extensions (**`.wav, .mp3, .avi,`** etc.) define the file types which your computer can play. Of course this function does require your computer to be able to handle sound or video if it is to work -- Windows uses the file extension to know how to play it.

Video files will require the free VLC Media Player to be installed (see http://www.vlcapp.com/).

Filename

The sound or video file-name might be

1. specified in a tag
2. the same name as the text file-name but with another extension such as **`.wav`**
3. found in the tag and interpreted using a table you have created previously. To do this, make each line like this:

```
<s1>=c:\my_corpus_sounds\angry_man.wav 560 2
<s2>=c:\my_corpus_sounds\happy_little_girl.mp3 980 2
```

where each line has the tag found in the text file, followed by **`=`** then the desired value.

If it is in the tag mark-up, to process a reference like **`<a href=http://gandalf.hit.uib.no/c/l/32401-1.mp3>`** in the source text, the **`=`** character is sufficient to define where the start of the filename begins. In this case, what follows **`=`** is a web address. For a text containing tags like this **`<sound$$C:\mysounds\talk.wav>`**, you'd put **`$$`** to show the start of filename. For the concordance example 212, **`soundfile=`** is adequate to identify where the filename begins.

The *media files folder* will be needed (for cases 1 and 2 above) if the sound files are not stored in the same folder as your text files.

How to play it

start-mark, duration-mark (optional)

You can indicate markers for start and duration if necessary. They would be needed if your tag contained e.g.

`<a href=http://gandalf.hit.uib.no/c/l/32401-1.mp3 start=0360 play=5>`

or if the sound file is the same as the text file with a different extension.

If so, you'd specify *duration-mark* as **`play=`** and *start-mark* as **`start=`** (because that is how they are marked in your text files)

Times are measured in seconds.

duration

You can specify a default duration as in the screenshot: 6 seconds. More may be needed especially if the sound tags are not spaced closely together in the text file.

If no start or duration indication is given, the whole sound or video file will be played.

If there are no duration and start position markers, the first number will be interpreted as start position and the second as duration, so a tag like this: **`<sound$$C:\mysounds\talk.wav 15 5>`** in your text file means "play **`c:\mysounds\talk.wav`** starting 15 seconds from the beginning and play for 5 seconds". If there's only one number as in **`<sound$$C:\mysounds\talk.wav 15>`**, that means "play **`c:\mysounds\talk.wav`** starting 15 seconds from the beginning and play for the default number of seconds".

defaults

The defaults are: play **`.mp3`** and **`.wav`** files. Once you've completed this, save your defaults 113 for next time.

See also: Sound and Video in Concord 212, Obtaining Sound and Video files 215, Overview of Tags 131, Making a Tag File 141, Tag Handling 132, Tag Concordancing 198, Showing Nearest Tags in Concord 199, Viewing the Tags 390, Types of Tag 145, Guide to handling the BNC

6.11 modify source texts

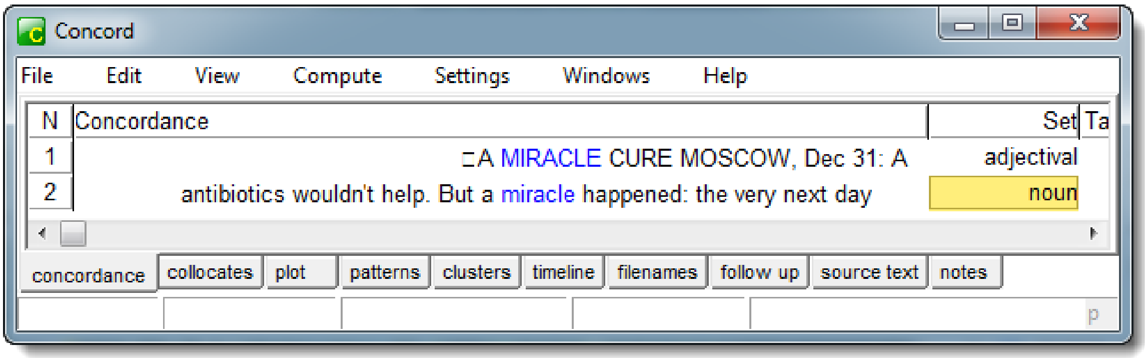
The point of it...

This function enables you to modify your original source texts as a result of concordance work you've done. In this way, your work can get saved in the source texts themselves. For example, you might want to save user-defined categories, or search-phrase results where you have decided a phrase is a multi-word unit.

Note: this procedure does alter your source texts. Before each is altered for the very first time, it is backed up (original filename with **`.original`** extension) but any change to your source texts or corpora must be done with caution!

User-defined categories

For example, suppose you have marked your concordance lines' Set column 168 like this:

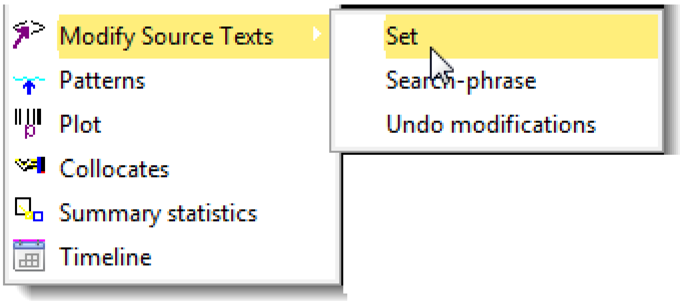


N	Concordance	Set	Ta
1	A MIRACLE CURE MOSCOW, Dec 31: A	adjectival	
2	antibiotics wouldn't help. But a miracle happened: the very next day	noun	

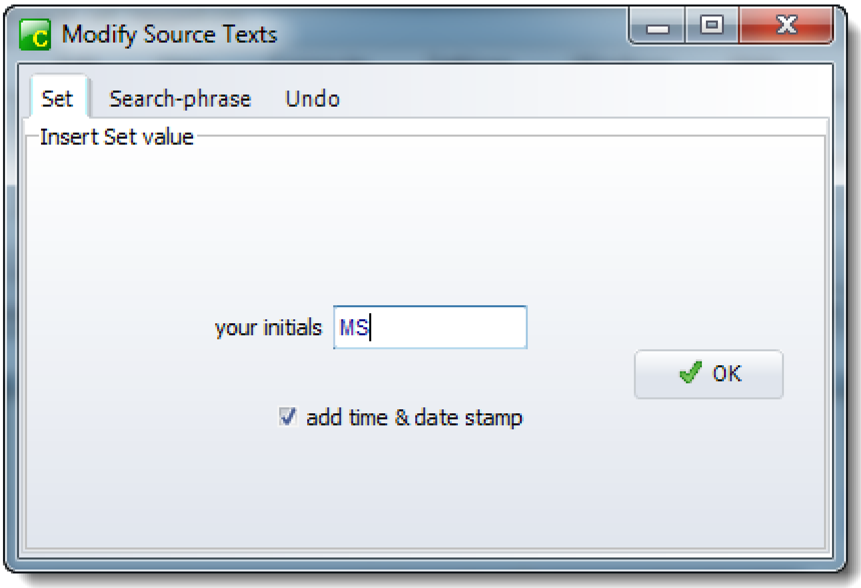
where the first line with **miracle** pre-modifies the noun **cure** and is marked **adjectival** but the second is an ordinary noun, and wish to save this in your original source text files.

How to do it

Choose *Compute* | *Modify Source Texts*.



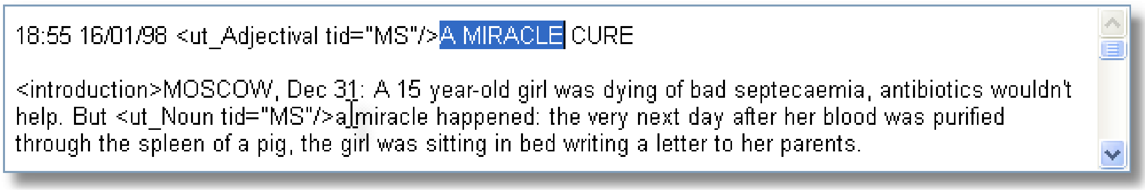
and if you want to save the Set choices, choose OK here:



your initials MS

☑ add time & date stamp

and the set choices will be marked as in this example:



18:55 16/01/98 <ut_Adjectival tid="MS"/>A MIRACLE CURE

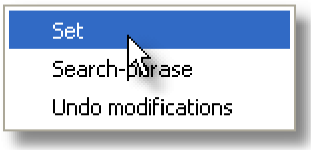
<introduction>MOSCOW, Dec 31: A 15 year-old girl was dying of bad septecaemia, antibiotics wouldn't help. But <ut_Noun tid="MS"/>a miracle happened: the very next day after her blood was purified through the spleen of a pig, the girl was sitting in bed writing a letter to her parents.

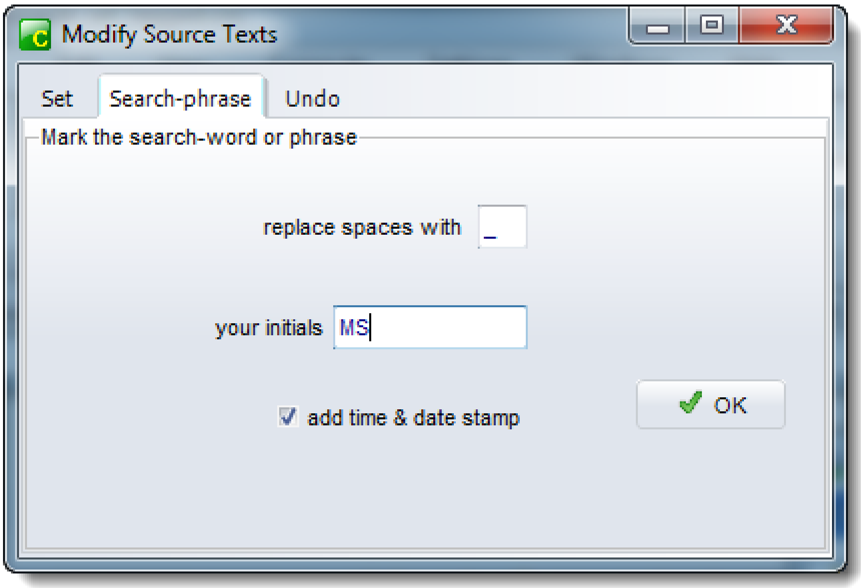
(seen by double-clicking the concordance line to show the source text).

Multi-word unit search phrase

Alternatively if you choose the search-phrase option:

and





Modify Source Texts

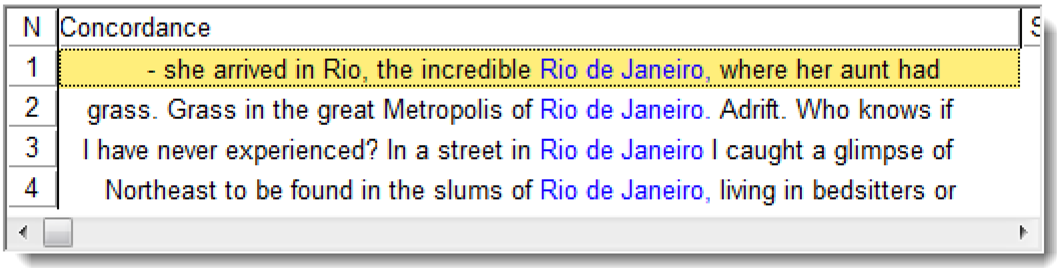
Mark the search-word or phrase

replace spaces with _

your initials MS

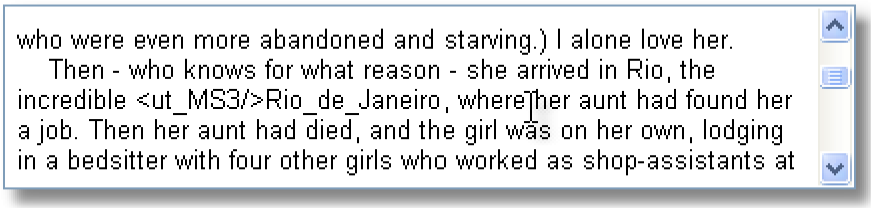
☑ add time & date stamp

then any search word containing a space will have underscores (or whatever other character you choose above) in it to establish multi-word units:



N	Concordance
1	- she arrived in Rio, the incredible Rio de Janeiro, where her aunt had
2	grass. Grass in the great Metropolis of Rio de Janeiro. Adrift. Who knows if
3	I have never experienced? In a street in Rio de Janeiro I caught a glimpse of
4	Northeast to be found in the slums of Rio de Janeiro, living in bedsitters or

Here, the search word or phrase was **`Rio de Janeiro`**, and the result of modifying the source texts was this:



who were even more abandoned and starving.) I alone love her.
Then - who knows for what reason - she arrived in Rio, the
incredible <ut_MS3/>Rio_de_Janeiro, where her aunt had found her
a job. Then her aunt had died, and the girl was on her own, lodging
in a bedsitter with four other girls who worked as shop-assistants at

Add Time & Date stamp option

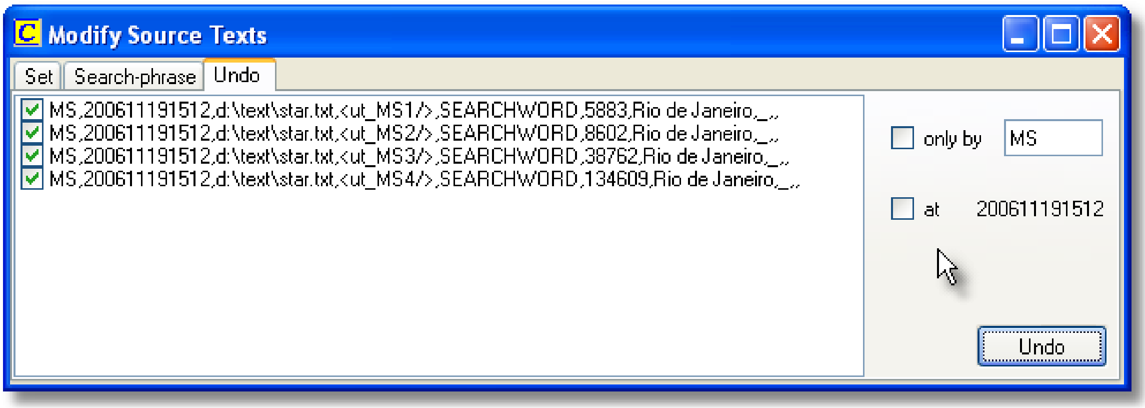
This keeps a log of all your changes, enabling the changes to be undone later.

Initials option

Adds your initials to the changes. Leave empty if not wanted. The **`<ut_MS3/>`** tag above means a user whose initials were **`MS`** made this change and it was the 3rd change.

To undo previous changes

If you have used the "time and date stamp" option shown above, you will be able to undo the modifications. The undo window shows all your log. You can choose all those done on a certain day, or by the person whose initials are visible at the right. Here we see the 4 modifications changing **Rio de Janeiro** into **Rio_de_Janeiro**.



☑ MS,200611191512,d:\text\star.txt,<ut_MS1/>,SEARCHWORD,5883,Rio de Janeiro,_,,
☑ MS,200611191512,d:\text\star.txt,<ut_MS2/>,SEARCHWORD,8602,Rio de Janeiro,_,,
☑ MS,200611191512,d:\text\star.txt,<ut_MS3/>,SEARCHWORD,38762,Rio de Janeiro,_,,
☑ MS,200611191512,d:\text\star.txt,<ut_MS4/>,SEARCHWORD,134609,Rio de Janeiro,_,,

□ only by MS

□ at 200611191512

See also: user-defined categories 168

6.12 XML text

What is XML?

XML text has angle-bracketed mark-up which provides additional information. For example the British National Corpus has text which is structured like this:

```
<s n="43">
<w c5="PNP" hw="i" pos="PRON">I </w><w c5="VVB" hw="mean"
pos="VERB">mean</w>
<c c5="PUN">, </c><w c5="AVQ" hw="where" pos="ADV">where </w>
<w c5="VDB" hw="do" pos="VERB">do </w><w c5="NN1-VVG" hw="eating"
pos="SUBST">eating </w>
<w c5="NN2" hw="disorder" pos="SUBST">disorders </w>
<w c5="VVB" hw="come" pos="VERB">come </w><w c5="PRP" hw="from"
pos="PREP">from</w>
<c c5="PUN">?</c>
</s>
```

<s> ... </s> signals a sentence
<w c5="PNP" hw="i" pos="PRON"> signals that the next word is a pronoun (coded **PNP**), head-word is **"i",**
<w c5="NN2" hw="disorder" pos="SUBST"> signals that the next word is a plural noun belonging to the head-word "**disorder"** and it's a substantive.

c5="NN2" is an ***attribute*** of the **<w** start-tag, **hw="disorder"** is another attribute. There can be many attributes in a start-tag. The **<c** start-tags have only one, but the **<w** start-tags have 3 in this BNC text.

WordSmith's handling of XML

By default, WordSmith simply ignores all the mark-up so a word list will only get the words in black inserted in it, a concordance will only see those words (**`I mean, where do eating disorders come from?`**).

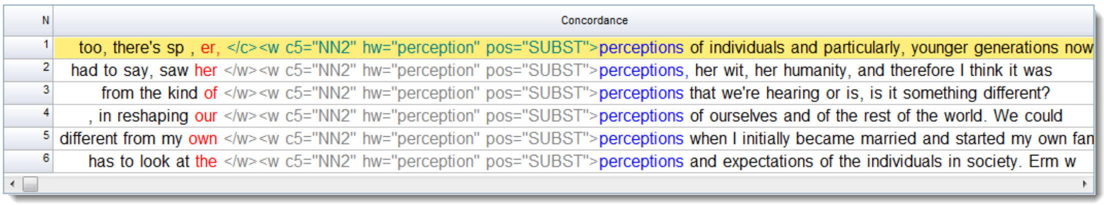
Searching using Attributes

If you want to search for all instances of NN2 forms (plural nouns), you'd need to type

`<w c5="NN2" * *>*`

as your search-word and answer yes to the question as to whether you're concordancing on tags.

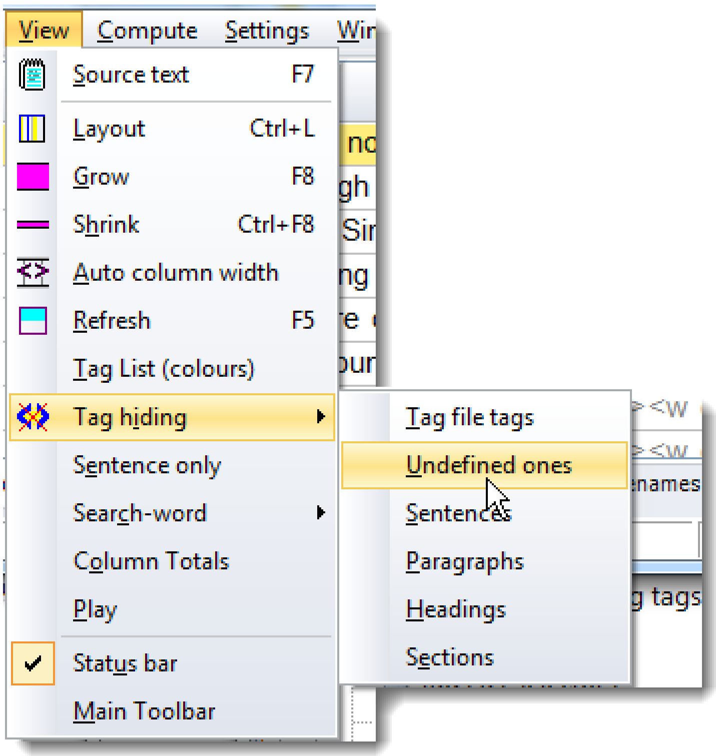
You would get results like this:

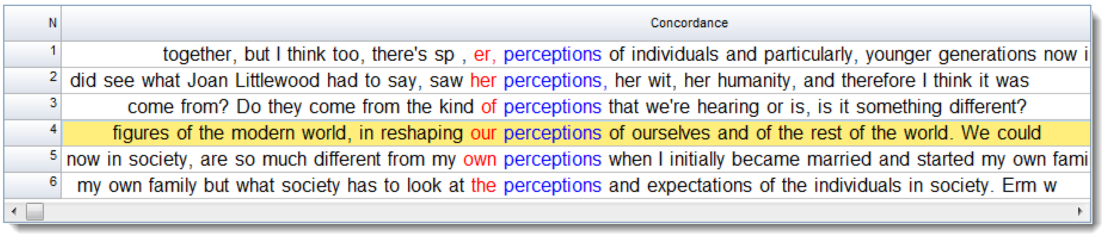


N	Concordance
1	too, there's sp , er, </c><w c5="NN2" hw="perception" pos="SUBST">perceptions of individuals and particularly, younger generations now
2	had to say, saw her </w><w c5="NN2" hw="perception" pos="SUBST">perceptions, her wit, her humanity, and therefore I think it was
3	from the kind of </w><w c5="NN2" hw="perception" pos="SUBST">perceptions that we're hearing or is, is it something different?
4	, in reshaping our </w><w c5="NN2" hw="perception" pos="SUBST">perceptions of ourselves and of the rest of the world. We could
5	different from my own </w><w c5="NN2" hw="perception" pos="SUBST">perceptions when I initially became married and started my own fan
6	has to look at the </w><w c5="NN2" hw="perception" pos="SUBST">perceptions and expectations of the individuals in society. Erm w

Hide the mark-up

If you prefer not to see all that the mark-up in grey, choose to hide the undefined mark-up





N	Concordance
1	together, but I think too, there's sp , er, perceptions of individuals and particularly, younger generations now i
2	did see what Joan Littlewood had to say, saw her perceptions, her wit, her humanity, and therefore I think it was
3	come from? Do they come from the kind of perceptions that we're hearing or is, is it something different?
4	figures of the modern world, in reshaping our perceptions of ourselves and of the rest of the world. We could
5	now in society, are so much different from my own perceptions when I initially became married and started my own fami
6	my own family but what society has to look at the perceptions and expectations of the individuals in society. Erm w

There is a button in the main tool which can show or hide mark-up, too.

Asterisks in your search-word

In the example above, we search on

`<w c5="NN2" * *>*`

`<w` because each start-tag where **`NN2`** forms are found starts with **`<w`** and the very first attribute is **`c5="NN2"`**. Then two asterisks to indicate that we aren't interested in the **`hw`** or **`pos`** attributes. Then a closing **>** and another asterisk because the word which follows will be right next to the **>** in our corpus.

For two successive parts of speech,

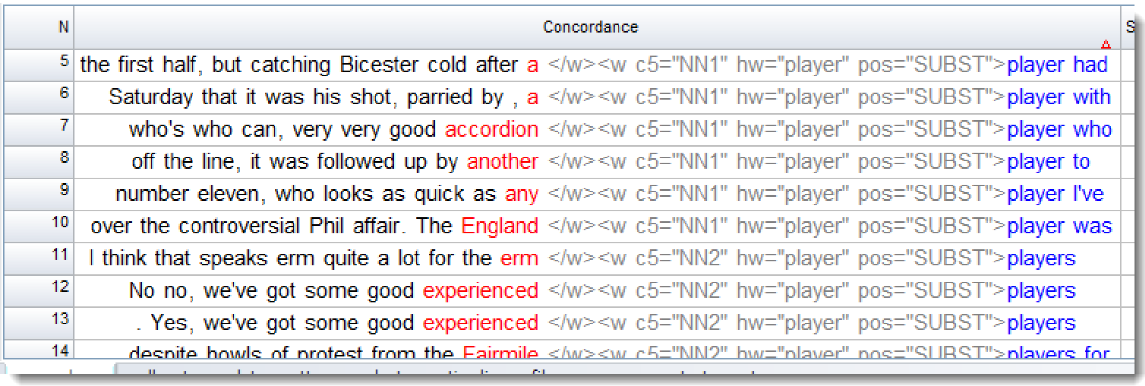
`<w c5="AT0" * *>* <w c5="NN1" * *>*`

looks for any article (the/a/an) followed by any singular count noun.

A search on

`<w c5="NN?" hw="player" *>*`

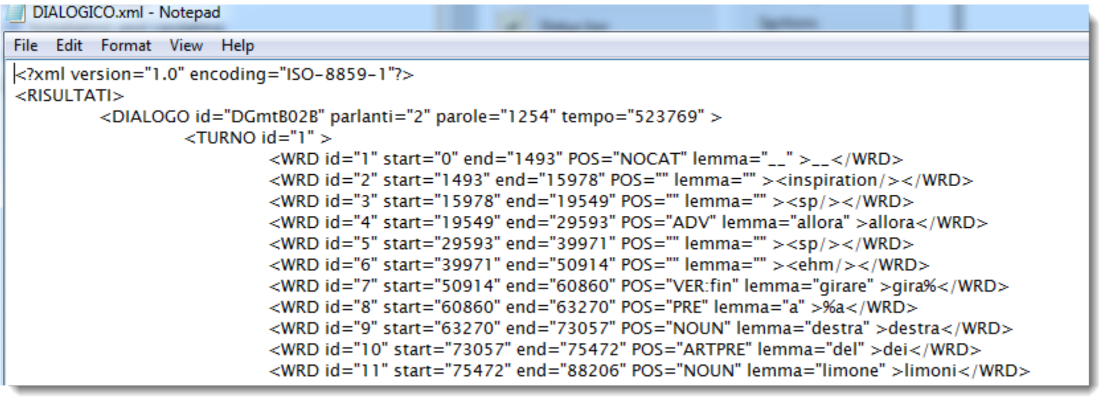
where we are allowing **`NN1`** or **`NN2`** and requiring the **`hw`** to be **`player`**,gets results like this:



N	Concordance	S
5	the first half, but catching Bicester cold after a </w><w c5="NN1" hw="player" pos="SUBST">player had	
6	Saturday that it was his shot, parried by , a </w><w c5="NN1" hw="player" pos="SUBST">player with	
7	who's who can, very very good accordion </w><w c5="NN1" hw="player" pos="SUBST">player who	
8	off the line, it was followed up by another </w><w c5="NN1" hw="player" pos="SUBST">player to	
9	number eleven, who looks as quick as any </w><w c5="NN1" hw="player" pos="SUBST">player I've	
10	over the controversial Phil affair. The England </w><w c5="NN1" hw="player" pos="SUBST">player was	
11	I think that speaks erm quite a lot for the erm </w><w c5="NN2" hw="player" pos="SUBST">players	
12	No no, we've got some good experienced </w><w c5="NN2" hw="player" pos="SUBST">players	
13	. Yes, we've got some good experienced </w><w c5="NN2" hw="player" pos="SUBST">players	
14	despite howls of protest from the Fairmile </w><w c5="NN2" hw="player" pos="SUBST">players for	

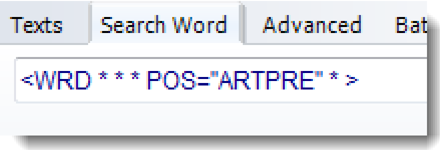
Another example

Searching Italian .XML containing text like this:



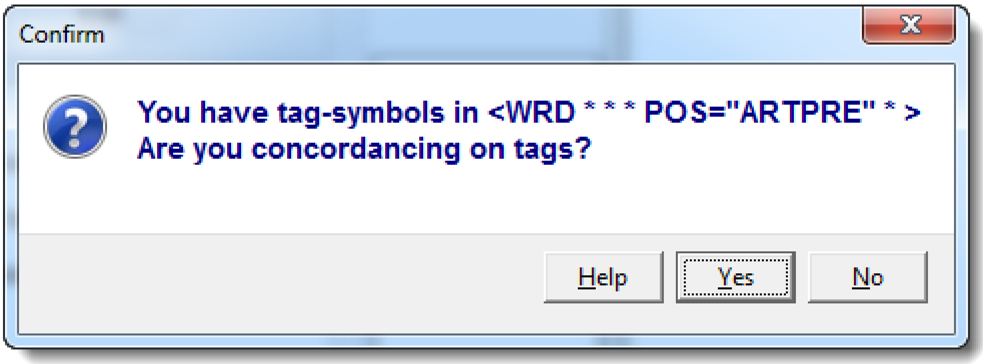
```
<?xml version="1.0" encoding="ISO-8859-1"?>
<RISULTATI>
          <DIALOGO id="DGmtB02B" parlanti="2" parole="1254" tempo="523769" >
                    <TURNO id="1" >
                              <WRD id="1" start="0" end="1493" POS="NOCAT" lemma="__" >__</WRD>
                              <WRD id="2" start="1493" end="15978" POS="" lemma="" ><inspiration/></WRD>
                              <WRD id="3" start="15978" end="19549" POS="" lemma="" ><sp/></WRD>
                              <WRD id="4" start="19549" end="29593" POS="ADV" lemma="allora" >allora</WRD>
                              <WRD id="5" start="29593" end="39971" POS="" lemma="" ><sp/></WRD>
                              <WRD id="6" start="39971" end="50914" POS="" lemma="" ><ehm/></WRD>
                              <WRD id="7" start="50914" end="60860" POS="VER:fin" lemma="girare" >gira%</WRD>
                              <WRD id="8" start="60860" end="63270" POS="PRE" lemma="a" >%a</WRD>
                              <WRD id="9" start="63270" end="73057" POS="NOUN" lemma="destra" >destra</WRD>
                              <WRD id="10" start="73057" end="75472" POS="ARTPRE" lemma="del" >dei</WRD>
                              <WRD id="11" start="75472" end="88206" POS="NOUN" lemma="limone" >limoni</WRD>
```

and wishing to find all cases of the ARTPRE part of speech, with the search-word specified like this



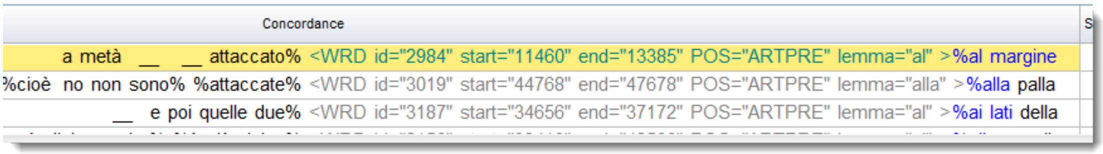
<WRD * * * POS="ARTPRE" * >

and answering yes to this:



You have tag-symbols in <WRD * * * POS="ARTPRE" * >
Are you concordancing on tags?

we get a considerable concordance with entries like this:



```
     a metà  __   __ attaccato% <WRD id="2984" start="11460" end="13385" POS="ARTPRE" lemma="al" >%al margine
%cioè  no non sono% %attaccate% <WRD id="3019" start="44768" end="47678" POS="ARTPRE" lemma="alla" >%alla palla
         __  e poi quelle due% <WRD id="3187" start="34656" end="37172" POS="ARTPRE" lemma="al" >%ai lati della
```

(I have no idea why there are % symbols in the source .XML, by the way.)

See also : Handling the BNC

WordSmith Tools Manual

Concord

Section VII

7 Concord

7.1 purpose

Concord is a program which makes a concordance 159 using plain text or web text files.

To use it you will specify a search word or phrase 159, which Concord will seek in all the text files you have chosen. It will then present a concordance display, and give you access to information about collocates of the search word, dispersion plots showing where the search word came in each file, cluster analyses showing repeated clusters of words (phrases) etc.

The point of it...

The point of a concordance is to be able to see lots of examples of a word or phrase, in their contexts. You get a much better idea of the use of a word by seeing lots of examples of it, and it's by seeing or hearing new words in context lots of times that you come to grasp the meaning of most of the words in your native language. It's by seeing the contexts that you get a better idea about how to use the new word yourself. A dictionary can tell you the meanings but it's not much good at showing you how to use the word.

Language students can use a concordancer to find out how to use a word or phrase, or to find out which other words belong with a word they want to use. For example, it's through using a concordancer that you could find out that in academic writing, a *paper* can *describe*, *claim*, or *show*, though it doesn't *believe* or *want* (**this paper wants to prove that* ...).

Language teachers can use the concordancer to find similar patterns so as to help their students. They can also use Concord to help produce vocabulary exercises, by choosing two or three search-words, blanking 168 them out, then printing 97.

Researchers can use a concordancer, for example when searching through a database of hospital accident records, to see whether *fracture* is associated with *fall*, *grease*, *ladder*. Or to examine historical documents to find all the references to *land ownership*.

Online step-by-step guide showing how

7.2 index

Explanations

What to do if it doesn't do what I want... 455
What is Concord and what's it for? 158
Collocation 179
Collocation Display 181
Plots 191

Settings

Procedures

see also : WordSmith Main Index 2

7.3 what is a concordance?

a set of examples of a given word or phrase, showing the context. A concordance of *give* might look like this:

```
... could not give me the time ...
... Rosemary, give me another ...
... would not give much for that ...
```

A concordancer searches through a text or a group of texts and then shows the concordance as output. This can be saved, printed, etc.

7.4 search-word or phrase

7.4.1 search word syntax

By default, Concord does a whole-word non-case-sensitive search.

Basic Examples

search word	finds
book	*Book* or *book* or *BoOk*

book*	*book, books, booking, booked*
***book**	*textbook* (but not *textbooks*)
b*	*banana, baby, brown etc.*
***ed**	*walked, wanted, picked etc.*
bo* in	*book in, books in, booking in* (but not *book into*)
book * hotel	*book a hotel, book the hotel, book my hotel*
bo* in*	*book in, books in, booking in, book into*
book?	*book, books, book; book.*
book^	*book, books*
b^^k	*book, back, bank*, etc.
==book==	*book* (but not *BOOK* or *Book*)
book/paperback	*book* or *paperback*

symbol	meaning	examples
*	disregard the end of the word, disregard a whole word	**tele*** ***ness** ***happi*** **book * hotel**
?	any single character (including punctuation) will match here	**Engl???** **?50.00**
#	any sequence of numbers, 0 to 9	**$#** **£#.00**
^	any single letter of the alphabet will match here	**Fr^nc^**
==	case sensitive	**==French==** **==Fr*==**
:\	means use a file for lots of search-words (see file-based search_words 161)	**c:\text\frd.txt**
/	separates alternative search-words. You can specify alternatives within an 80-character overall limit	**may/can/will**
<>	beginning & end of tags	**<w NN1>**

Advanced Search-word Syntax

If you want to use ***, ? , == , #, ^ , :\, >, <** or **/** as a character in your search word, put it in double quotes. Examples:

```
"*"
Why"?"
and"/"or
":\"
"<"
```

Don't forget that question-marks come at the end of words (in English anyway) so you might need **`*"?"`**

If you need to type in a symbol using a code, they can be symbolised like this: **`{CHR(xxx)}`** where **`xxx`** is the decimal number of the code. Examples: **`{CHR(13)}`** is a carriage-return, **`{CHR(10)}`** is a line-feed, **`{CHR(9)}`** is a tab. To represent **`<Enter>`** which comes at the end of paragraphs and sometimes at the end of each line, you'd type **`{CHR(13)}{CHR(10)}`** which is carriage-return followed immediately by line-feed.

`{CHR(34)}` refers to double inverted commas.

`{CHR(46)}` is a full stop. There is a list of codes at http://www.asciitable.com/

You can also use hex format for numbers, e.g. **`#x9`** for tab, **`#x22`** for double inverted commas.

Tags

You can also specify tags in your search-word if your text is tagged.

Examples:

symbol	meaning	examples
<w NN1>*	single common noun (BNC World)	***book, chair, elephant***
<w c5="NN1" * *>*	single common noun (BNC XML edition)	***book, chair, elephant***
<w NN?>*	singular or plural common noun	***book, chairs***
<w NN1>t*	any single noun beginning with **`T`** or **`t`**	***table, teacher***
<w NN1>* <w NN1>*	two single common nouns in sequence	***campaign manager***

for XML formats see XML text handling 153

See also: Tag Concordancing 198, Context Word 202, Modify source texts 149, Ignore punctuation 225, Wildcards 28

7.4.2 file-based search-words

The point of it...

To save time typing in complex searches.

You may want to do a standard search repeatedly on different sub-corpora.

Or as Concord allows an almost unlimited number of entries, you may wish to do a concordance involving many search-words or phrases 159.

The space for typing in multiple search-words is limited to 80 characters (including / etc.). If your preferred search-words will exceed this limit or you wish to use a standardised search, you can prepare a file containing all the search-words.

How to do it...

A sample (**`Documents\wsmith6\concordance_search_words.txt`**) is included with the distribution files.

Use a Windows editor (e.g. Notepad) to prepare your own. Each one must be on a separate line of your file. No comment lines can be included, though blank lines may be inserted for readability.

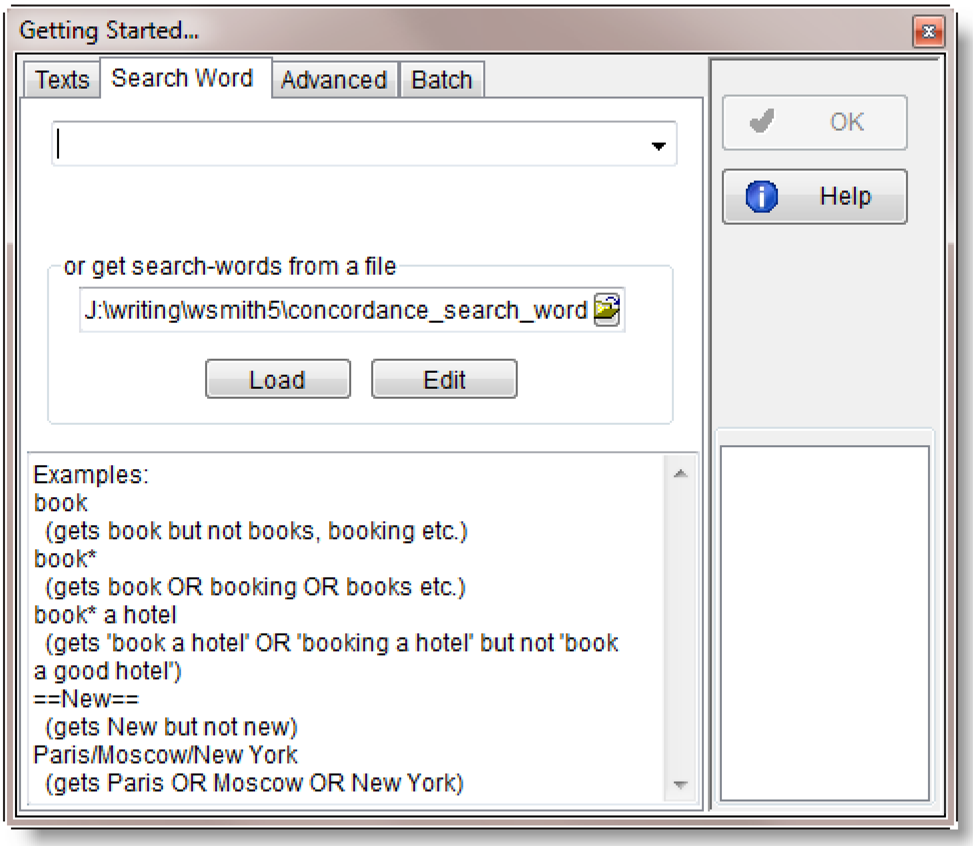
If you want to require a context for a given word, put **`context:=`** as in this example:

```
 book context:=hotel
```

(which seeks **book** and only shows results if **`hotel`** comes in the context horizons).

Then, instead of typing in each word or phrase in the Search Word dialogue box, just browse for the file.

Then press *Load* to read the entries (or *Clear* if you change your mind).



J:\writing\wsmith5\concordance_search_word

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==
 (gets New but not new)
Paris/Moscow/New York
 (gets Paris OR Moscow OR New York)

Lemmas and file-based concordancing

Note that where Concord has been called up[438] from WordList, and the highlighted word in the word list is the head entry with lemmas[270], a temporary file will be created, listing the whole set of lemmas, and Concord will use this file-based search-word procedure to compute the concordance. The temporary file will be stored in your **`Documents\wsmith6`** folder unless you're running on a network in which case it'll be in Windows' temporary folder, e.g. **`\windows\temp`**. It's up to you to delete the temporary file.

Automated file-based concordances

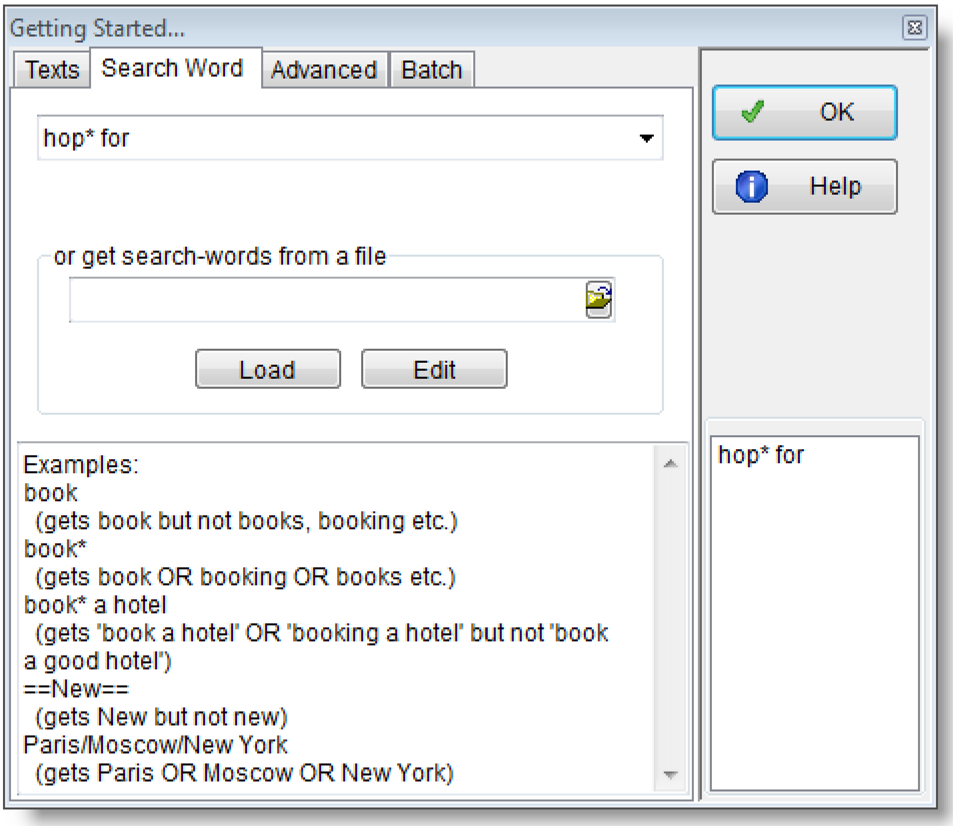
If you want Concord to process a whole lot of different search-words, saving each result as it goes along so you can get a lot of work done with WordSmith unattended, choose *SW Batch* under

Concordance Settings 163.

7.4.3 search-word and other settings

Search Word or Phrase and/or Tag

Type the word or phrase 159 Concord will search for when making the concordance, or (below) the name of a file of search words 161. You may also choose from a history list 435 of your previous search words. For details of syntax, see Search Word Syntax 159 or the set of examples shown in this screenshot:



hop* for

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==
(gets New but not new)
Paris/Moscow/New York
(gets Paris OR Moscow OR New York)

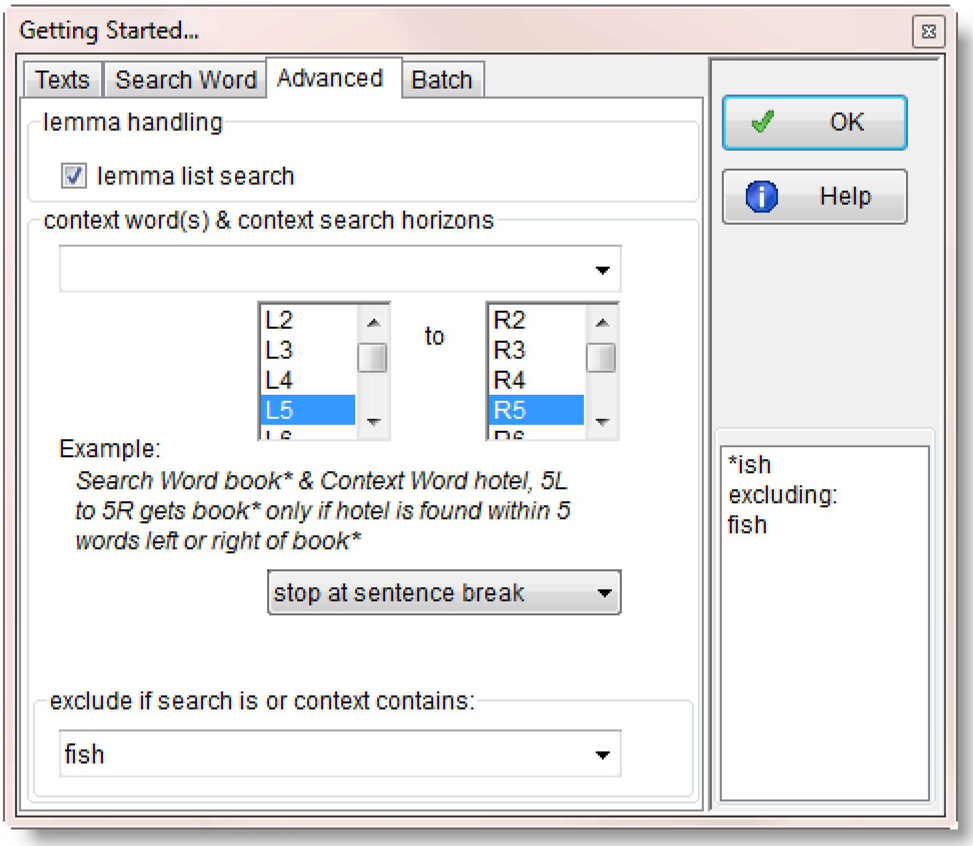
hop* for

If you want to do many concordances in a file-based search 161, first prepare a small text file containing the search words, e.g. containing

```
this
that
the other
==Major*==
```

Press the file button to locate your text file, the press the *Load* button. This will then change its name to something like *Clear 4*, where 4 means as in the example above that there are 4 different search-words to be concordanced. See "Batch" below for details on saving each one under a separate filename, otherwise all the searches will be combined into the same concordance.

Advanced searches



☑ lemma list search

context word(s) & context search horizons: L5 to R5

Example:
Search Word book & Context Word hotel, 5L to 5R gets book* only if hotel is found within 5 words left or right of book**

stop at sentence break

exclude if search is or context contains: fish

*ish
excluding:
fish

lemma list search

This option requires you to have chosen and loaded a lemma file 274. If the lemma file you've loaded specifies for example **speak -> speaks, spoke, spoken** then if your search-word is **speak**, the concordance will contain examples of all four forms.

Context word(s) and search horizons

You may wish to find a word or phrase depending on the context. In that case you can specify context word(s) which you want, or which you do not want (and if found will mean that entry is not used).

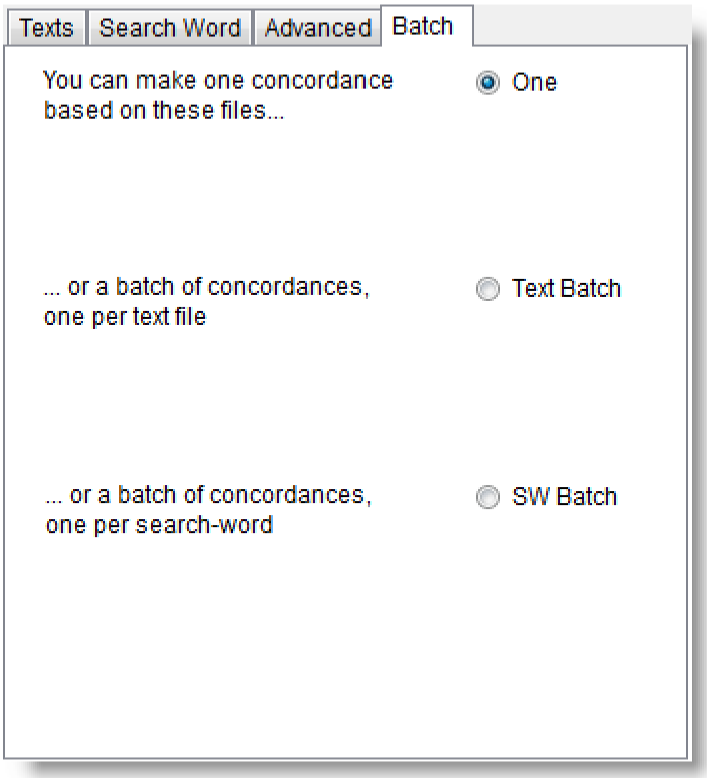
For example, if the search word is **book*** and the context word is **hotel**, you'll get **book, books, booked, booking, bookable**, but only if **hotel** is found within your Context Search Horizons 202. Or if the search word is **book*** and the *exclude if* box has **hotel**, you'll get **book, books, booked, booking, bookable**, as long as **hotel** is *not* found within your context search horizons. Or if the search word is ***ish** and the exclusion specifies **fish**, you'll get **yellowish, greenish**, etc. but not **fish**.

You may type tag mark-up in here too, e.g. search for **book** with a context word <**ADJ>*** in position up to L3 will find book with a preceding adjective if your text has that sort of mark-up and if you've defined a tag file including **<ADJ>.**

In the screenshot above you see that "stop at sentence break" has been selected, meaning that a collocation search will only go left or right of the search-word up to a sentence-end. This is further explained here 188.

Batch

Suppose you're concordancing book* in 20 text files: you might want One concordance based on all 20 files (the default), or instead 20 separate concordances in a zipped batch 39 which can be viewed separately (Text Batch). If you have multiple search-words in a file-based search 161 as explained above, you may want each result saved separately (SW *Batch*).



You can make one concordance based on these files... ● One

... or a batch of concordances, one per text file ○ Text Batch

... or a batch of concordances, one per search-word ○ SW Batch

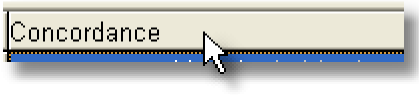
Other settings affecting a concordance are available too:
see WordSmith Controller Concordance Settings 222; Typing characters 420, Accented characters 419; Choosing Language 81, Context Word(s) & Context Search Horizons 202

7.5 advice

You have a listing showing all the concordance lines in a window. You can scroll up and down and left or right with the mouse or with the cursor keys.

Sort the lines

If you have a lot of lines you should certainly sort them. A concordance is read vertically, not horizontally. You are looking for repeated patterns, such as the presence of lots of the same sorts of words to the right or left of your search-word. Click the bar at the top to start a sort.



The Columns

These show the details for each entry: the entry number, the concordance line, set, tag, word-position (e.g. 1st word in the text is 1), paragraph and sentence position, source text file name, and how far into the file it comes (as a percentage). See below for an explanation of the **purple blobs**. The statistics are computed automatically based on the language settings.

Set

This is where you can classify the entries yourself, using any letter, into user-defined categories 168. Supposing you want to sort out verb uses from noun uses, you can press V or N. To type more (eg. "Noun"), double-click the entry in the set column and type what you want. If you have more than one search-word 159, you will find the Set column filled with the search-word for each entry. To clear the current entry, you can type the number 0. To clear the whole Set column, choose Edit | Clear Set column.

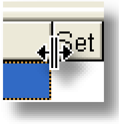
Tag

This column shows the tag context 199.

More context?

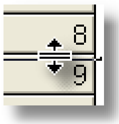
Stretching the display to see more

You can pull the concordance display to widen its column. Just place the mouse cursor on the bar between one column and another; when the cursor changes shape you can pull the whole column.

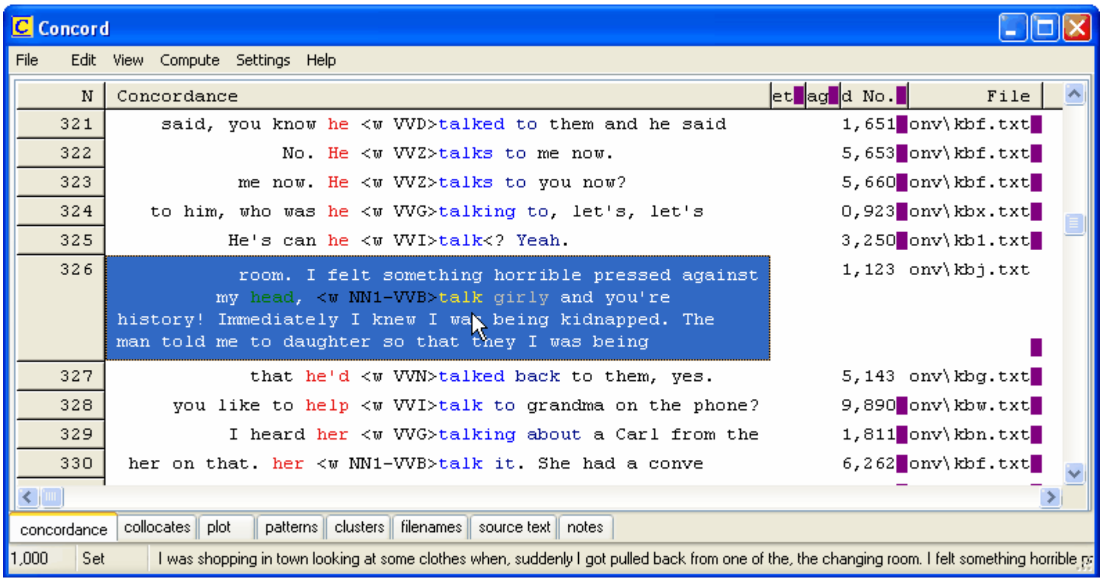


Stretch one line to see more context

The same applies to each individual row: place the mouse cursor between one row and another in the grey numbered area, and drag.



Or press ■ (F8) to "grow" all the rows, or ▬ (Ctrl+F8) to shrink them. Or press the numeric key-pad 8 to grow the current line as shown below. (Use numeric key-pad 2 to shrink it.)


Concord
File Edit View Compute Settings Help
concordance collocates plot patterns clusters filenames source text notes

N	Concordance	et▮	ag▮	d No.▮	File
321	said, you know he <w VVD>talked to them and he said			1,651▮	onv\kbf.txt▮
322	No. He <w VVZ>talks to me now.			5,653▮	onv\kbf.txt▮
323	me now. He <w VVZ>talks to you now?			5,660▮	onv\kbf.txt▮
324	to him, who was he <w VVG>talking to, let's, let's			0,923▮	onv\kbx.txt▮
325	He's can he <w VVI>talk<? Yeah.			3,250▮	onv\kb1.txt▮
326	room. I felt something horrible pressed against my head, <w NN1-VVB>talk girly and you're history! Immediately I knew I was being kidnapped. The man told me to daughter so that they I was being			1,123	onv\kbj.txt ▮
327	that he'd <w VVN>talked back to them, yes.			5,143	onv\kbg.txt▮
328	you like to help <w VVI>talk to grandma on the phone?			9,890▮	onv\kbw.txt▮
329	I heard her <w VVG>talking about a Carl from the			1,811▮	onv\kbn.txt▮
330	her on that. her <w NN1-VVB>talk it. She had a conve			6,262▮	onv\kbf.txt▮

1,000 Set I was shopping in town looking at some clothes when, suddenly I got pulled back from one of the, the changing room. I felt something horrible p

Viewing the original text-file

(if it is still on the disk where it was when the concordance was originally created)

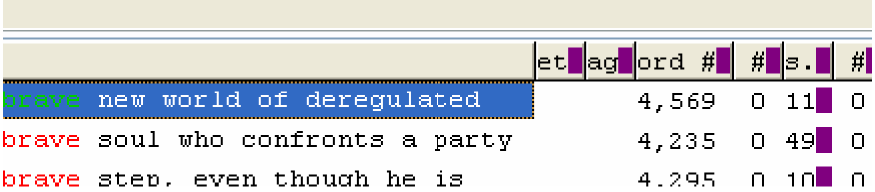
Double-click the concordance column, and the source text window will load the file and highlight the search word 116.

Or double-click the filename column, it will open in Notepad for editing.

Other things you may wonder about

Weird purple marks

In the screenshot you will see purple marks where any column is not wide enough to show all the data. The reason is that numbers are often not fully visible and you might otherwise get the wrong impression. For example in the concordance below, the *Word #* column shows **4,569** but the true number might be **14,569**. Pull the column wider and the purple lines disappear.



	et▮	ag▮	ord #▮	#▮	s.▮	#▮
brave new world of deregulated			4,569	0	11▮	0
brave soul who confronts a party			4,235	0	49▮	0
brave step, even though he is			4,295	0	10▮	0

Status bar

The status bar 449 panels show

- the number of entries (1,000 in the "stretch one line" screenshot above)
- whether we're in "Set" or "Edit" mode;
- the current concordance line from its start.

See also:

Re-sorting 208 your concordance lines
Follow-up 171 searches
User-defined categories 168
Altering the View 220
Blanking out 168 the search-word

Padding the search-word with spaces (use the search-word padding menu item to put a space on each side of the search-word)
Collocation 179 (words in the neighbourhood of the search-word)
Plot 191 (plots where the search-word came in the texts)
Clusters 175 (groups of words in your concordance)
Text segments in Concord 218
Editing the concordance 206
Time-lines 126
Zapping entries 206
Saving and printing 211
Window Management 127

7.6 blanking

In a concordance, to blank out the search-words with asterisks, just press the spacebar (or choose *View* | *Blanked out*). Press it again to restore them.

The point of it...

A blanked-out concordance is useful when you want to create an exercise. This one has *give* and *put* mingled:

```
   ... could not ********** me the time ...
   ... Rosemary, ********** me another ...
   ... would not ********** much for that ...
   ... could not ********** up with him ...
   ... so you'll ********** him a present ...
   ... will soon ********** up smoking ...
   ... he should ********** it over here ...
```

Concord will give equal space to the blanks so that the size of the blank doesn't give the game away.

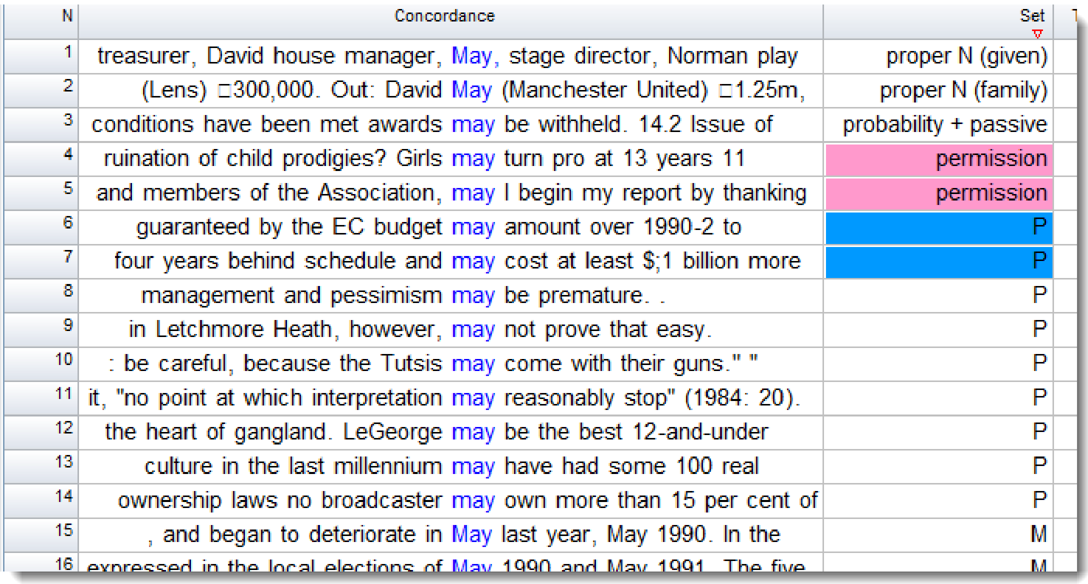
See also : Other main Controller settings for Concord 222

7.7 Category or Set

7.7.1 set column categories

The point of it...

You may want to classify entries in your own way, e.g. separating adjectival uses from nominal ones, or sorting according to different meanings.



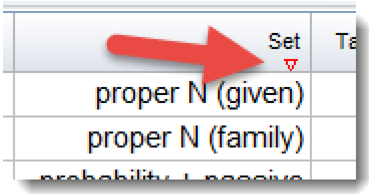
N	Concordance	Set
1	treasurer, David house manager, May, stage director, Norman play	proper N (given)
2	(Lens) □300,000. Out: David May (Manchester United) □1.25m,	proper N (family)
3	conditions have been met awards may be withheld. 14.2 Issue of	probability + passive
4	ruination of child prodigies? Girls may turn pro at 13 years 11	permission
5	and members of the Association, may I begin my report by thanking	permission
6	guaranteed by the EC budget may amount over 1990-2 to	P
7	four years behind schedule and may cost at least $;1 billion more	P
8	management and pessimism may be premature. .	P
9	in Letchmore Heath, however, may not prove that easy.	P
10	: be careful, because the Tutsis may come with their guns." "	P
11	it, "no point at which interpretation may reasonably stop" (1984: 20).	P
12	the heart of gangland. LeGeorge may be the best 12-and-under	P
13	culture in the last millennium may have had some 100 real	P
14	ownership laws no broadcaster may own more than 15 per cent of	P
15	, and began to deteriorate in May last year, May 1990. In the	M
16	expressed in the local elections of May 1990 and May 1991. The five	M

Here the user has used **P** where **may** has to do with probability and **M** if it's a month. In addition, some items have been labelled in more detail. You may wish also to modify your original texts 149 to include this annotation work you've done.

How to do it

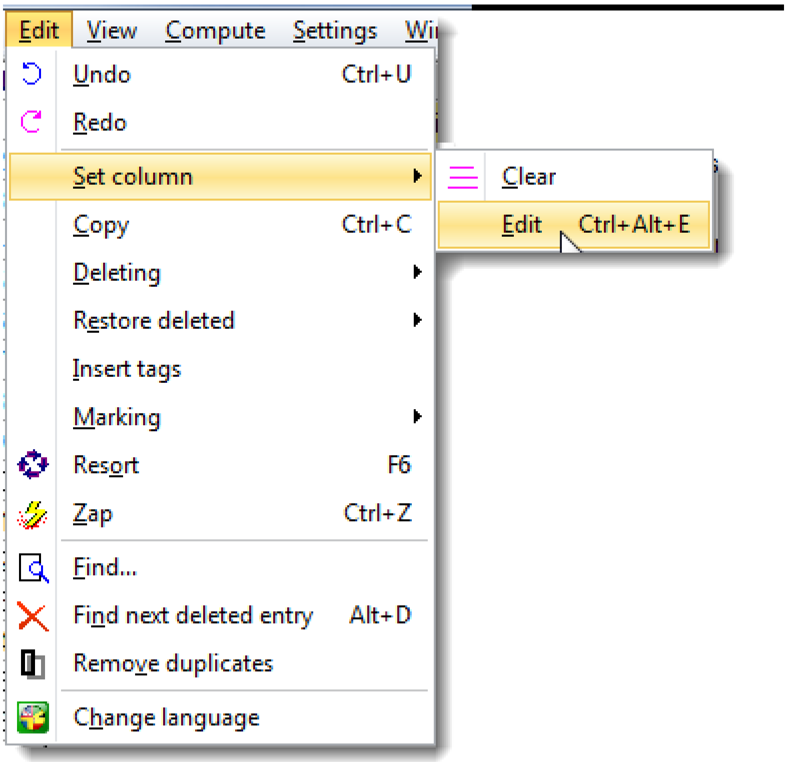
If you simply press a letter or number key while the edit v. set v. type-in 428 mode is on Set you will get the concordance line marked with that letter or number in the Set column.

You can sort 208 the concordance lines using these categories, simply by clicking on the header Set, which will have a small triangle showing it's sorted.



Set	Ta
proper N (given)	
proper N (family)	
probability + passive	

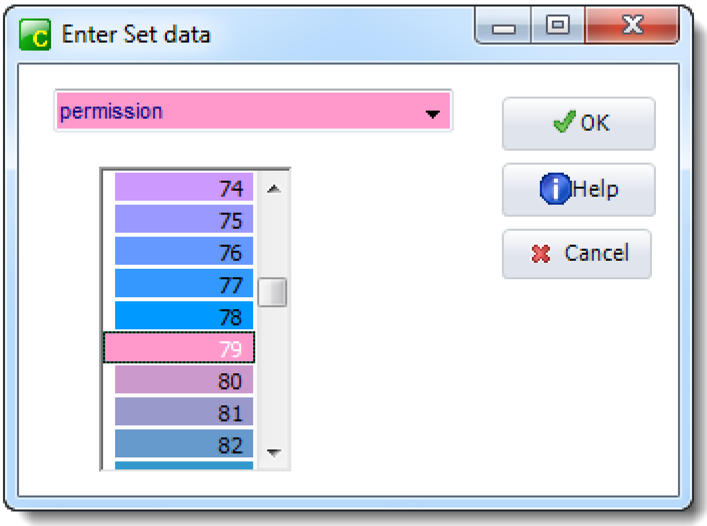
To enter the same value for various rows, first select the rows or mark 112 them, then choose *Set column | Edit*



then type in a suitable value.

Colours

If you want to type something longer and optionally in a specific colour, double-click the set column and you'll get a chance to type more.



Here the word **permission** has been typed and the colour 79 has been dragged onto the box.

Clearing the Set column

To correct a mistake, press the zero key; that will remove any text and colour from the selected entry. (If you press the spacebar you will get blanking 168.)

See also : Colour categories 51, modify your source texts 149, edit v. type-in 428 mode.

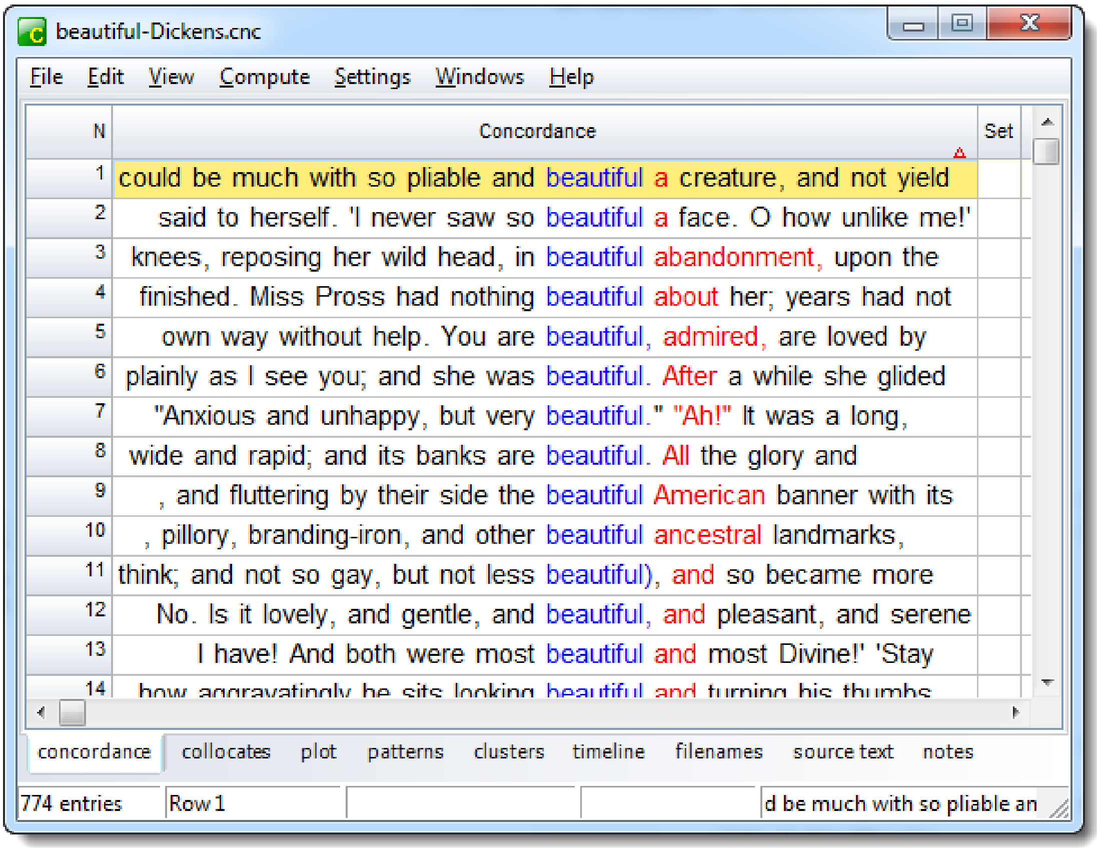
7.7.2 colour categories

The point of it...

The idea is to follow up a large concordance by breaking it down into specific sub-sections, so one can see how many of each sub-type are found in the whole list.

Example

The screen-shot below came from a concordance of **beautiful** in Charles Dickens:



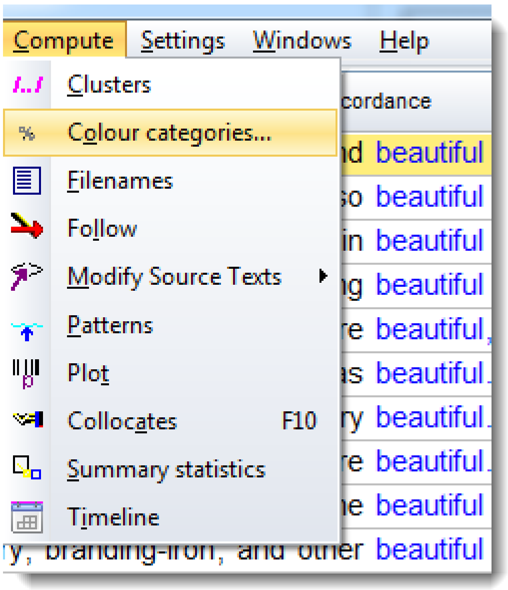
N	Concordance	Set
1	could be much with so pliable and beautiful a creature, and not yield	
2	said to herself. 'I never saw so beautiful a face. O how unlike me!'	
3	knees, reposing her wild head, in beautiful abandonment, upon the	
4	finished. Miss Pross had nothing beautiful about her; years had not	
5	own way without help. You are beautiful, admired, are loved by	
6	plainly as I see you; and she was beautiful. After a while she glided	
7	"Anxious and unhappy, but very beautiful." "Ah!" It was a long,	
8	wide and rapid; and its banks are beautiful. All the glory and	
9	, and fluttering by their side the beautiful American banner with its	
10	, pillory, branding-iron, and other beautiful ancestral landmarks,	
11	think; and not so gay, but not less beautiful), and so became more	
12	No. Is it lovely, and gentle, and beautiful, and pleasant, and serene	
13	I have! And both were most beautiful and most Divine!' 'Stay	
14	how aggravatingly he sits looking beautiful and turning his thumbs	

774 entries | Row 1 | d be much with so pliable an

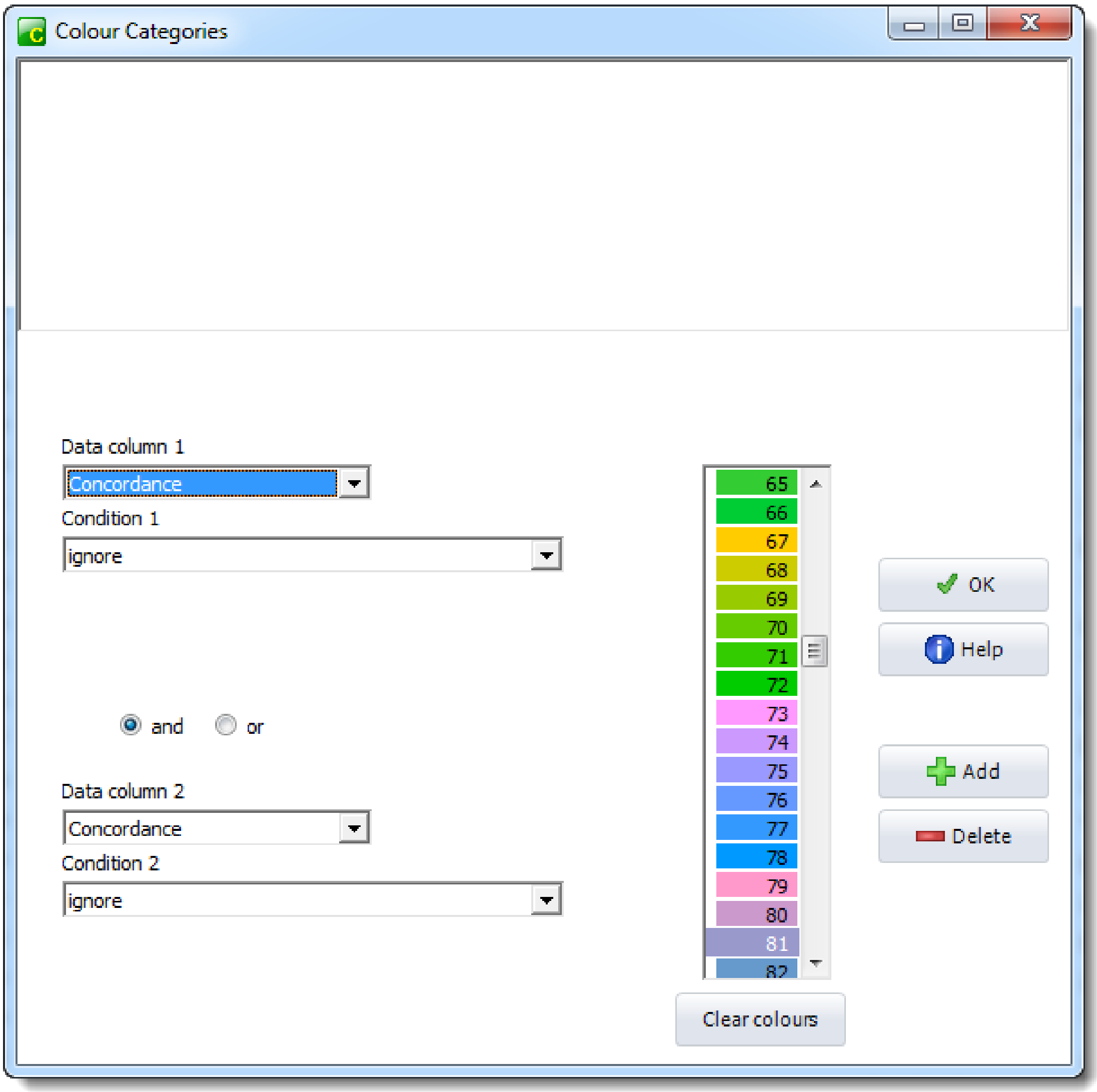
There are 774 lines. Looking through them, it became apparent that Dickens was fond of the collocations **beautiful creature** and **beautiful face**, but how many are of **beautiful creature** or similar (such as **so.. beautiful a creature** in line 1) and what proportion of the lines is that?

How to do it

Choose *Compute* | *Colour Categories* in the menu.



which opens up the colour categories box:



Colour Categories

Data column 1: Concordance

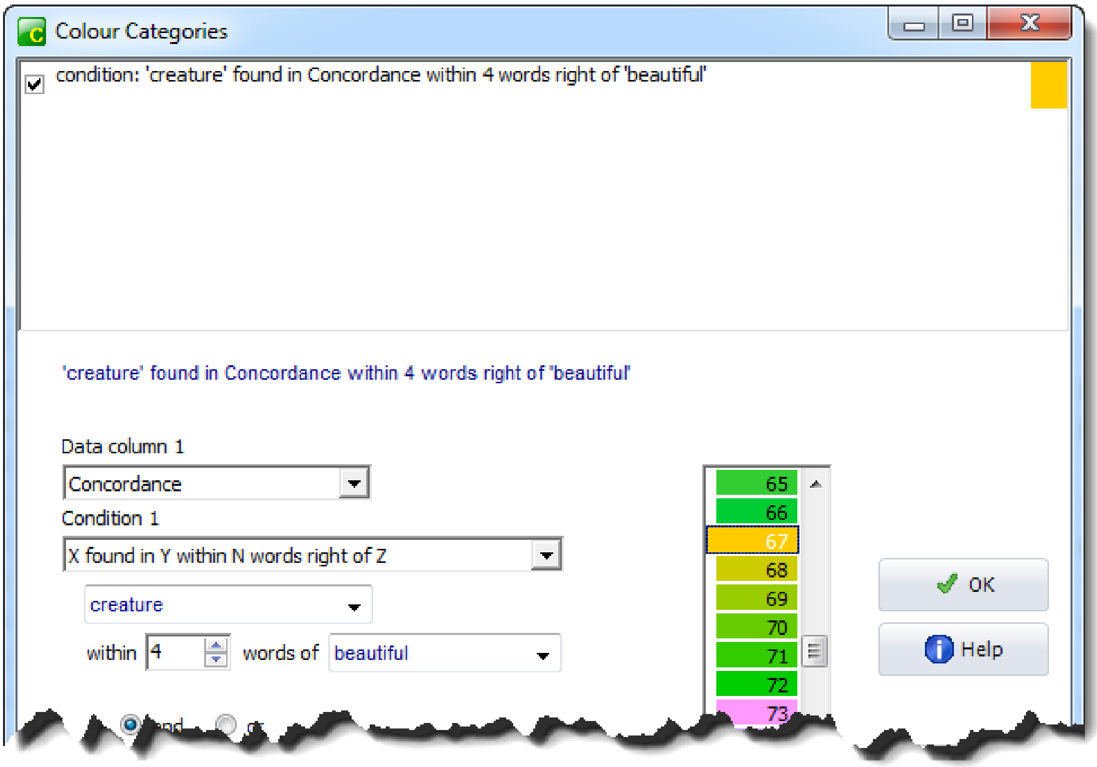
Condition 1: ignore

● and ○ or

Data column 2: Concordance

Condition 2: ignore

Here we have completed the spaces so as to get cases of beautiful ... with creature up to 4 words away to the right, and chosen



☑ condition: 'creature' found in Concordance within 4 words right of 'beautiful'

'creature' found in Concordance within 4 words right of 'beautiful'

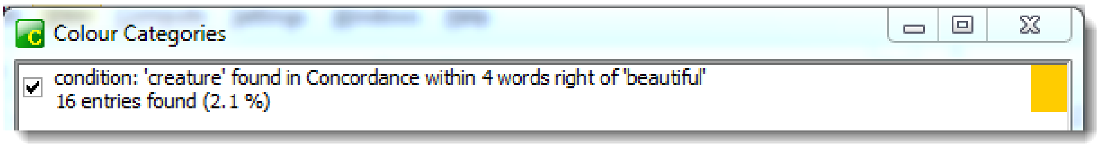
Data column 1: Concordance

Condition 1: X found in Y within N words right of Z

creature

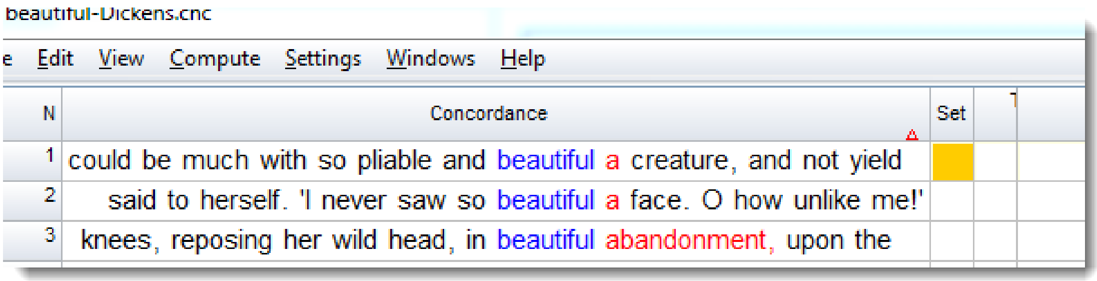
within 4 words of beautiful

to colour yellow any which meet this condition. On pressing OK we find out there are 16, representing just over 2% of the lines.



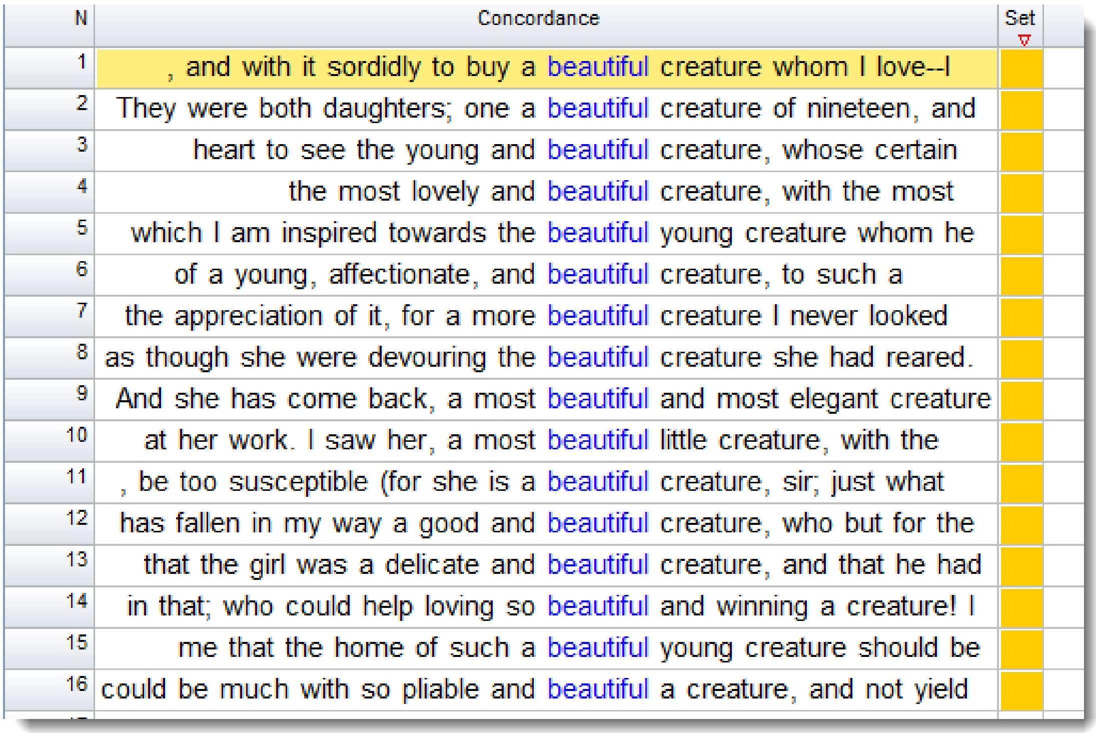
☑ condition: 'creature' found in Concordance within 4 words right of 'beautiful'
16 entries found (2.1 %)

and looking at the concordance the first line is now marked:



N	Concordance	Set
1	could be much with so pliable and beautiful a creature, and not yield	🟨
2	said to herself. 'I never saw so beautiful a face. O how unlike me!'	
3	knees, reposing her wild head, in beautiful abandonment, upon the	

Where are the other 15? To find them, simply sort on the Set column.



N	Concordance	Set ▽
1	, and with it sordidly to buy a beautiful creature whom I love--I	🟨
2	They were both daughters; one a beautiful creature of nineteen, and	🟨
3	heart to see the young and beautiful creature, whose certain	🟨
4	the most lovely and beautiful creature, with the most	🟨
5	which I am inspired towards the beautiful young creature whom he	🟨
6	of a young, affectionate, and beautiful creature, to such a	🟨
7	the appreciation of it, for a more beautiful creature I never looked	🟨
8	as though she were devouring the beautiful creature she had reared.	🟨
9	And she has come back, a most beautiful and most elegant creature	🟨
10	at her work. I saw her, a most beautiful little creature, with the	🟨
11	, be too susceptible (for she is a beautiful creature, sir; just what	🟨
12	has fallen in my way a good and beautiful creature, who but for the	🟨
13	that the girl was a delicate and beautiful creature, and that he had	🟨
14	in that; who could help loving so beautiful and winning a creature! I	🟨
15	me that the home of such a beautiful young creature should be	🟨
16	could be much with so pliable and beautiful a creature, and not yield	🟨

This function applies to word lists and other data too, and is explained in more detail the main colour categories 51 section. The set column itself can contain characters or words as well as colours, as explained in the set column 168 section.

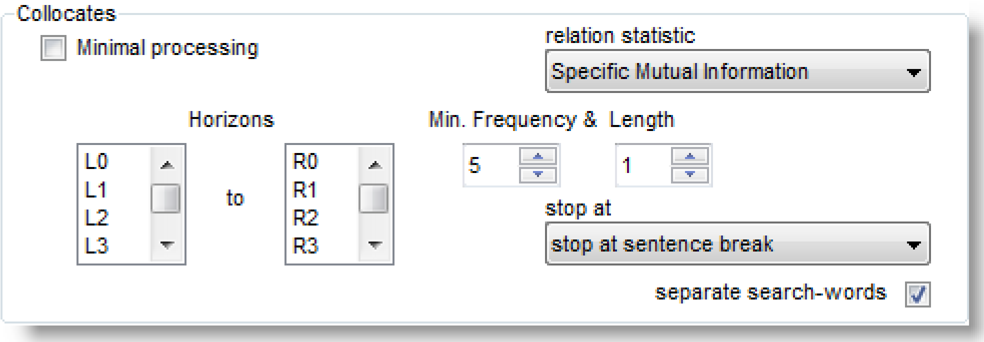
7.8 clusters

The point of it...

These word clusters help you to see patterns of repeated phraseology in your concordance, especially if you have a concordance with several thousand lines. Naturally, they will usually contain the search-word itself, since they are based on concordance lines.
Another feature in **Concord** which helps you see patterns is Patterns 207.

How it does it...

Clusters are computed automatically if this is not disabled in the main Controller 222 settings for Concord (*Concord Settings*) where you will see something like this:



Collocates
□ Minimal processing
relation statistic: Specific Mutual Information
Horizons: L0 L1 L2 L3 to R0 R1 R2 R3
Min. Frequency & Length: 5 1
stop at: stop at sentence break
separate search-words ☑

where your usual default settings are controlled. "Minimal processing", if checked, means do not compute collocates, clusters, patterns etc. when computing a concordance. (They can always be computed later if the source text files are still present.)

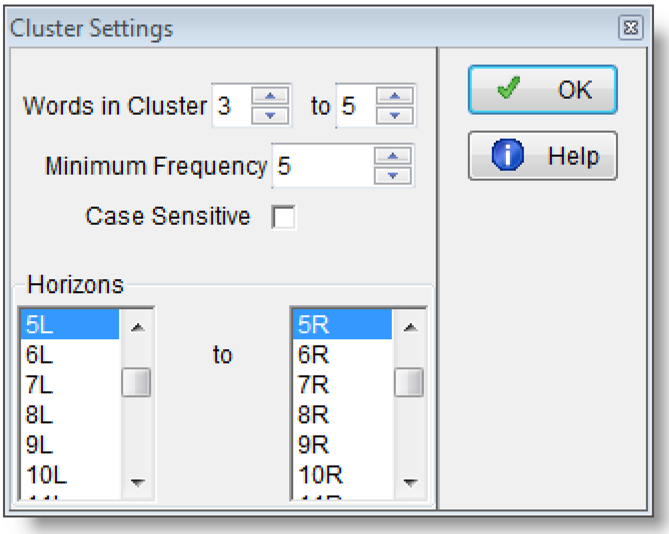
Clusters are sought within these limits: default: 5 words left and right of the search word, but up to 25 left and 25 right allowed. The default is for clusters to be three words in length and you can choose how many of each must be found for the results to be worth displaying (say 3 as a minimum frequency).

Clusters are calculated using the existing concordance lines. That is, any line which has not been deleted or zapped is used for computing clusters.

As with WordList index clusters 278, the idea of "stop at sentence breaks 188" (there are other alternatives) is that a cluster which spans across two sentences is not likely to make sense.

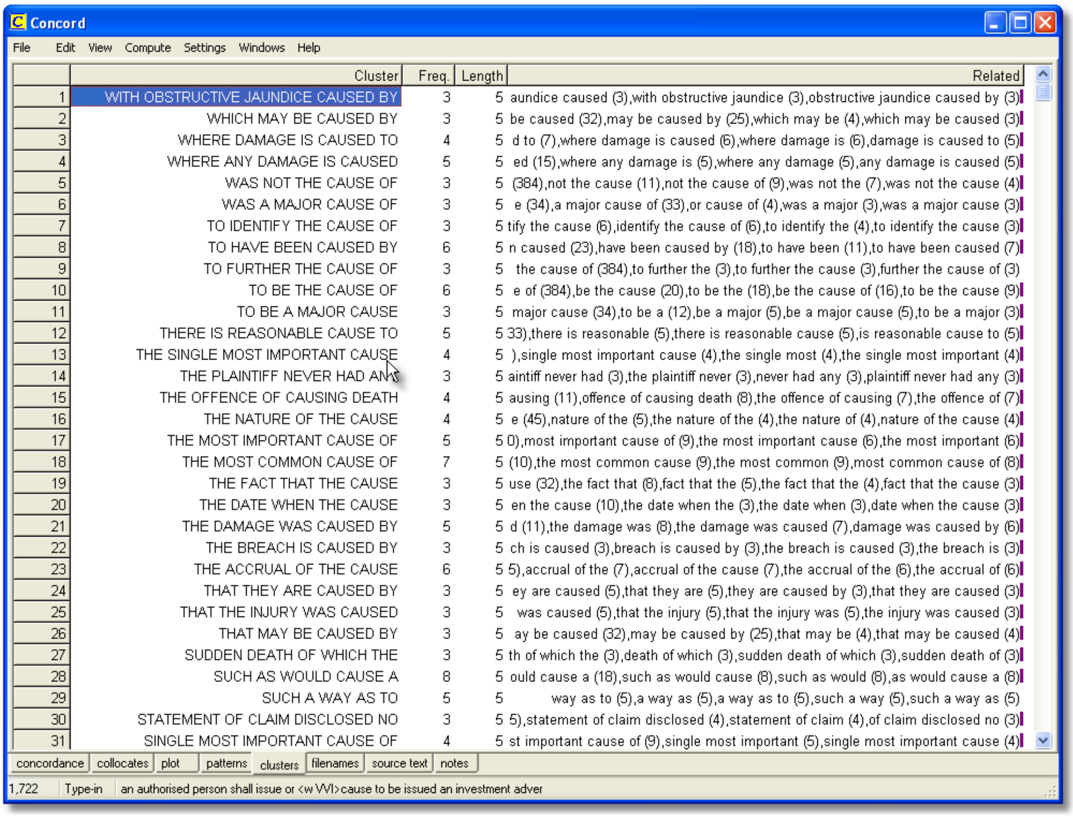
Re-computing clusters (/../)

The default clusters computed may not suit, (and you may want to recompute after deleting some lines), so you can also choose *Compute* | *Clusters* (/../) in the Concord menu, so as to choose how many words a cluster should have (cluster size 2 to 4 words recommended), and alter the other settings.



Cluster Settings
Words in Cluster 3 to 5
Minimum Frequency 5
Case Sensitive □
Horizons: 5L 6L 7L 8L 9L 10L to 5R 6R 7R 8R 9R 10R

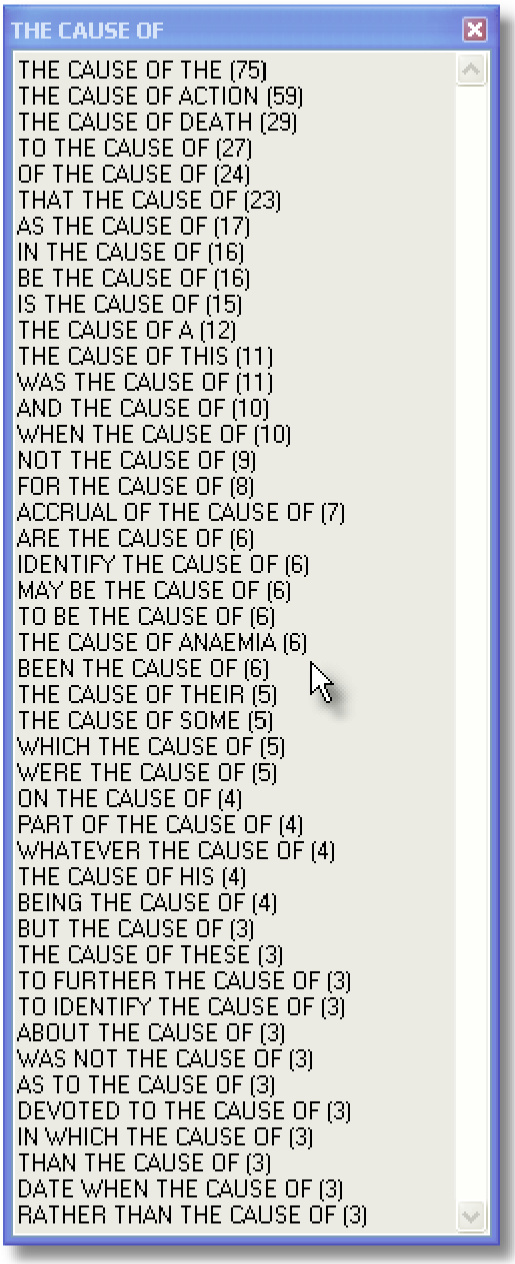
When you press OK, clusters will be computed. In this case we have asked for 3- to 5-word clusters and get results like this:



	Cluster	Freq.	Length	Related
1	WITH OBSTRUCTIVE JAUNDICE CAUSED BY	3	5	aundice caused (3),with obstructive jaundice (3),obstructive jaundice caused by (3)
2	WHICH MAY BE CAUSED BY	3	5	be caused (32),may be caused by (25),which may be (4),which may be caused (3)
3	WHERE DAMAGE IS CAUSED TO	4	5	d to (7),where damage is caused (6),where damage is (6),damage is caused to (5)
4	WHERE ANY DAMAGE IS CAUSED	5	5	ed (15),where any damage is (5),where any damage (5),any damage is caused (5)
5	WAS NOT THE CAUSE OF	3	5	(384),not the cause (11),not the cause of (9),was not the (7),was not the cause (4)
6	WAS A MAJOR CAUSE OF	3	5	e (34),a major cause of (33),or cause of (4),was a major (3),was a major cause (3)
7	TO IDENTIFY THE CAUSE OF	3	5	tify the cause (6),identify the cause of (6),to identify the (4),to identify the cause (3)
8	TO HAVE BEEN CAUSED BY	6	5	n caused (23),have been caused by (18),to have been (11),to have been caused (7)
9	TO FURTHER THE CAUSE OF	3	5	the cause of (384),to further the (3),to further the cause (3),further the cause of (3)
10	TO BE THE CAUSE OF	6	5	e of (384),be the cause (20),to be the (18),be the cause of (16),to be the cause (9)
11	TO BE A MAJOR CAUSE	3	5	major cause (34),to be a (12),be a major (5),be a major cause (5),to be a major (3)
12	THERE IS REASONABLE CAUSE TO	5	5	33),there is reasonable (5),there is reasonable cause (5),is reasonable cause to (5)
13	THE SINGLE MOST IMPORTANT CAUSE	4	5	),single most important cause (4),the single most (4),the single most important (4)
14	THE PLAINTIFF NEVER HAD ANY	3	5	aintiff never had (3),the plaintiff never (3),never had any (3),plaintiff never had any (3)
15	THE OFFENCE OF CAUSING DEATH	4	5	ausing (11),offence of causing death (8),the offence of causing (7),the offence of (7)
16	THE NATURE OF THE CAUSE	4	5	e (45),nature of the (5),the nature of the (4),the nature of (4),nature of the cause (4)
17	THE MOST IMPORTANT CAUSE OF	5	5	0),most important cause of (9),the most important cause (6),the most important (6)
18	THE MOST COMMON CAUSE OF	7	5	(10),the most common cause (9),the most common (9),most common cause of (8)
19	THE FACT THAT THE CAUSE	3	5	use (32),the fact that (8),fact that the (5),the fact that the (4),fact that the cause (3)
20	THE DATE WHEN THE CAUSE	3	5	en the cause (10),the date when the (3),the date when (3),date when the cause (3)
21	THE DAMAGE WAS CAUSED BY	5	5	d (11),the damage was (8),the damage was caused (7),damage was caused by (6)
22	THE BREACH IS CAUSED BY	3	5	ch is caused (3),breach is caused by (3),the breach is caused (3),the breach is (3)
23	THE ACCRUAL OF THE CAUSE	6	5	5),accrual of the (7),accrual of the cause (7),the accrual of the (6),the accrual of (6)
24	THAT THEY ARE CAUSED BY	3	5	ey are caused (5),that they are (5),they are caused by (3),that they are caused (3)
25	THAT THE INJURY WAS CAUSED	3	5	was caused (5),that the injury (5),that the injury was (5),the injury was caused (3)
26	THAT MAY BE CAUSED BY	3	5	ay be caused (32),may be caused by (25),that may be (4),that may be caused (4)
27	SUDDEN DEATH OF WHICH THE	3	5	th of which the (3),death of which (3),sudden death of which (3),sudden death of (3)
28	SUCH AS WOULD CAUSE A	8	5	ould cause a (18),such as would cause (8),such as would (8),as would cause a (8)
29	SUCH A WAY AS TO	5	5	way as to (5),a way as (5),a way as to (5),such a way (5),such a way as (5)
30	STATEMENT OF CLAIM DISCLOSED NO	3	5	5),statement of claim disclosed (4),statement of claim (4),of claim disclosed no (3)
31	SINGLE MOST IMPORTANT CAUSE OF	4	5	st important cause of (9),single most important (5),single most important cause (4)

1,722 Type-in an authorised person shall issue or <w VVI>cause to be issued an investment adver

The clusters have been sorted on the Length column so as to bring the 5-word clusters to the top. At the right there is a set of "Related" clusters, and for most of these lines it is impossible to see all of their entries. To solve this problem, double-click any line in the Related column and another window opens. Here is the window showing what clusters are related to the 3-word cluster, **`the cause of`**, which is the most frequent cluster in this set:



THE CAUSE OF

THE CAUSE OF THE (75)
THE CAUSE OF ACTION (59)
THE CAUSE OF DEATH (29)
TO THE CAUSE OF (27)
OF THE CAUSE OF (24)
THAT THE CAUSE OF (23)
AS THE CAUSE OF (17)
IN THE CAUSE OF (16)
BE THE CAUSE OF (16)
IS THE CAUSE OF (15)
THE CAUSE OF A (12)
THE CAUSE OF THIS (11)
WAS THE CAUSE OF (11)
AND THE CAUSE OF (10)
WHEN THE CAUSE OF (10)
NOT THE CAUSE OF (9)
FOR THE CAUSE OF (8)
ACCRUAL OF THE CAUSE OF (7)
ARE THE CAUSE OF (6)
IDENTIFY THE CAUSE OF (6)
MAY BE THE CAUSE OF (6)
TO BE THE CAUSE OF (6)
THE CAUSE OF ANAEMIA (6)
BEEN THE CAUSE OF (6)
THE CAUSE OF THEIR (5)
THE CAUSE OF SOME (5)
WHICH THE CAUSE OF (5)
WERE THE CAUSE OF (5)
ON THE CAUSE OF (4)
PART OF THE CAUSE OF (4)
WHATEVER THE CAUSE OF (4)
THE CAUSE OF HIS (4)
BEING THE CAUSE OF (4)
BUT THE CAUSE OF (3)
THE CAUSE OF THESE (3)
TO FURTHER THE CAUSE OF (3)
TO IDENTIFY THE CAUSE OF (3)
ABOUT THE CAUSE OF (3)
WAS NOT THE CAUSE OF (3)
AS TO THE CAUSE OF (3)
DEVOTED TO THE CAUSE OF (3)
IN WHICH THE CAUSE OF (3)
THAN THE CAUSE OF (3)
DATE WHEN THE CAUSE OF (3)
RATHER THAN THE CAUSE OF (3)

"Related" clusters are those which overlap to some extent with others, so that **`the cause of`** overlaps with **`devoted to the cause of`**, etc. The procedure seeks out cases where the whole of a cluster is found within another cluster.

See also: general information on clusters 448, WordList Clusters 278, Word Clouds 128.

7.9 Collocation

7.9.1 what is collocation?

What's a "collocate"?

Collocates are the words which occur in the neighbourhood of your search word. Collocates of *letter* might include *post, stamp, envelope*, etc. However, very common words like *the* will also collocate with *letter*.

and "colligation"?

Linkages between neighbouring words which involve grammatical items are often referred to as *colligation*. That **rely** is typically followed by a preposition in English is a colligational fact.

The point of it...

The point of all this is to work out characteristic lexical patterns by finding out which "friends" words typically hang out with. It can be hard to see overall trends in your concordance lines, especially if there are lots of them. By examining collocations in this way you can see common lexical and grammatical patterns of co-occurrence.

Options

You may compute a concordance with or without collocates (*minimal processing* 226): without is slightly quicker and will take up less room on your hard disk. The default is to compute with collocates.
The number of collocates stored will depend on the collocation horizons 180.
You can re-compute collocates after editing your concordance.
If you want to filter your collocate list, use a match list 92 or stop list 120.
Re-sort 189 a collocate list in a variety of ways.
You can see the strength of relationship 289 between the word and the search-word which the concordance was based on.

Collocates can be viewed 181 after the concordance has been computed.

Technical Note

The literature 417 on collocation has never distinguished very satisfactorily between collocates which we think of as "associated" with a word (letter - stamp) on the one hand, and on the other, the words which do actually co-occur with the word (letter - my, this, a, etc.).
We could call the first type "coherence collocates" and the second "neighbourhood collocates" or "horizon collocates". It has been suggested that to detect coherence collocates is very tricky, as once we start looking beyond a horizon of about 4 or 5 words on either side, we get so many words that there is more noise than signal in the system.

KeyWords allows you to study Associates 241, which are a pointer to "coherence collocates". **Concord** will supply "neighbourhood collocates". **WordList** and **Concord** allow you also to study relationships between words 289.

See also: collocation display 181, collocation settings 187, collocation relationship 180, relationships between words display 289.

7.9.2 collocate horizons

The collocate horizons represent the number of collocates Concord will find to the left and right of your search word, and the distance used by **KeyWords** in searching out plot-links 247. The default 113 is 5L,5R (5 to left and 5 to right) but you can go up to 25 on either side. You can set whether to set collocation boundaries such as sentence, paragraph breaks 188 too.

To set collocation horizons and other **Concord** settings, choose Concord Main Controller Settings 222, or in the main **WordSmith** Controller 4, choose *Concord Settings*.

See also: Collocate Settings 187

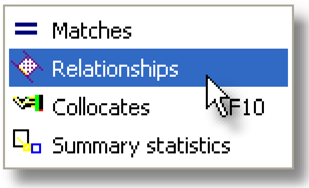
7.9.3 collocation relationship

The point of it...

The idea is to find out *how strongly* each collocate relates to the search-word near which it was found. MI (or other relevant statistic 289) is not computed by default for a collocate list.

How to compute it

In the Concord menu, choose *Compute* | *Relationships:*



Steps

1. Suppose you have made a concordance using all the files in **`Documents\wsmith6\text\shakespeare`** and have done a concordance on *love*. You see collocates such as *Romeo, hate, the, Juliet, Nurse* etc. All these show a "Relation" score of "??" because they haven't yet been computed.
2. If you haven't done so yet, use WordList to make a word list of the same text files (or if you prefer, use some other reference corpus 447). Make sure the reference corpus 447 file is what you prefer.
3. Now choose the menu item ✥ and Concord will use the reference corpus filename. It will look up each of your collocates in the word list and compute MI using the information in the reference corpus word list.

You can choose a different statistic in the main Controller Concord settings 222.

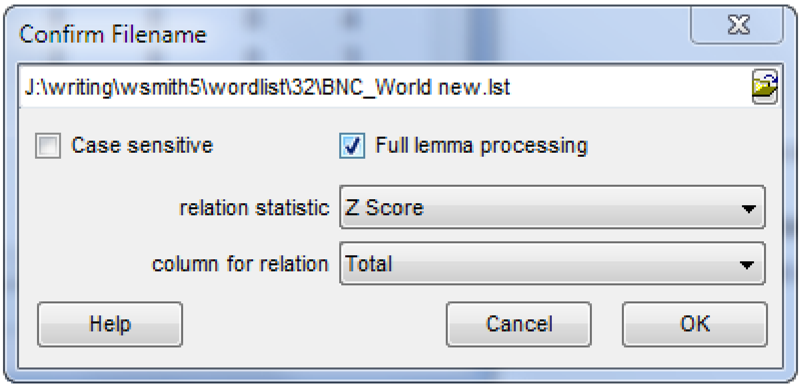
Note: the procedure goes through your collocates and tries to find each in the word-list. If absent, you get a blank result. If one of your search-terms has a space in it such as *Friar Lawrence*, an ordinary single-word word list won't know its frequency and you will be asked to supply it. If you don't know, you should compute a concordance on that search-phrase over the same corpus first.

Full lemma processing, case sensitive

These are only relevant if your word list has any lemmatised 270 entries, or it is a case-sensitive word list and you wish processing to respect case-sensitivity.

Relation statistic

Choose which type of relation you wish to compute. The default is Specific Mutual Information but in the screenshot Z score has been chosen.



Confirm Filename

J:\writing\wsmith5\wordlist\32\BNC_World new.lst

□ Case sensitive ☑ Full lemma processing

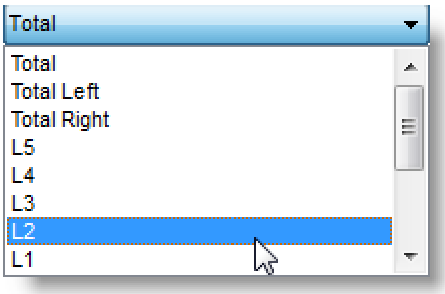
relation statistic: Z Score

column for relation: Total

Column for relation

The default is "Total". If you choose Total you're computing the relationship across the current collocation horizons 180 set.

If you prefer to examine the relationship at only one position instead, you may:



Total

See also: Collocation 179, Collocate display 181, Mutual Information 289

7.9.4 collocates display

Display

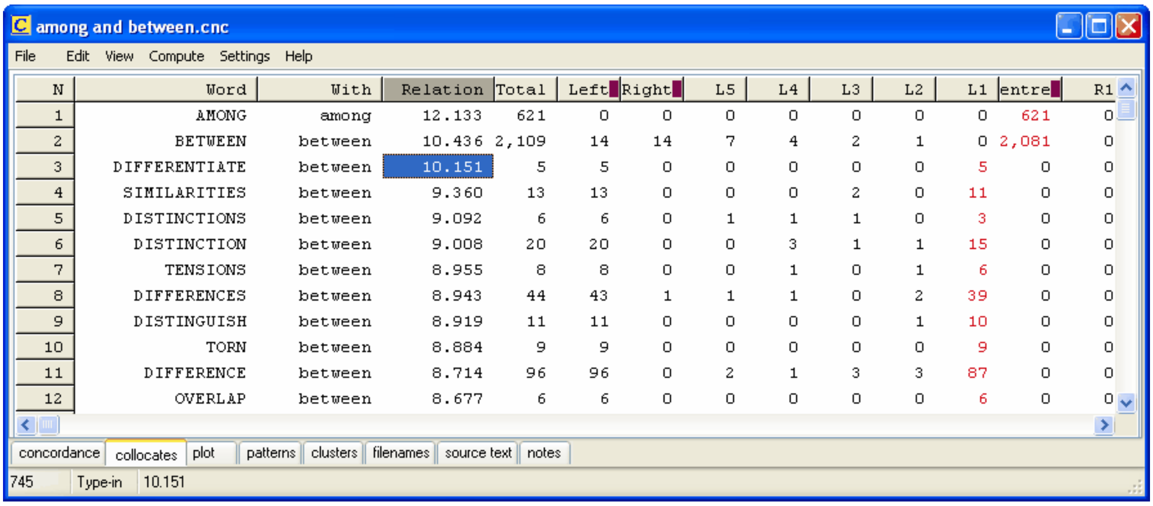
The collocation display initially shows the collocates in frequency order.

Beside each word and the search-word which the concordance was based on, you'll see the

strength of relationship 289 between the two (or 0.000 if it hasn't yet been computed).
Then, the total number of times it co-occurred with the search word in your concordance, and a total for Left and Right of the search-word. Then a detailed break-down, showing how many times it cropped up 5 words to the left, 4 words to the left, and so on up to 5 words to the right. The centre position (where the search word came) is shown with an asterisk.

The number of words to left and right depends on the collocation horizons 180.
The numbers are:
  the total number of times the word was found in the neighbourhood of the search word
  the total number of times it came to the left of the search-word
  the total number of times it came to the right of the search-word
  a set of individual frequencies to the left of the search word (5L, i.e. 5 words to the left, 4L .. 1L)
  a Centre column, representing the search-word
  a set of individual frequencies to the right of the search word (1R, 2R, etc.)

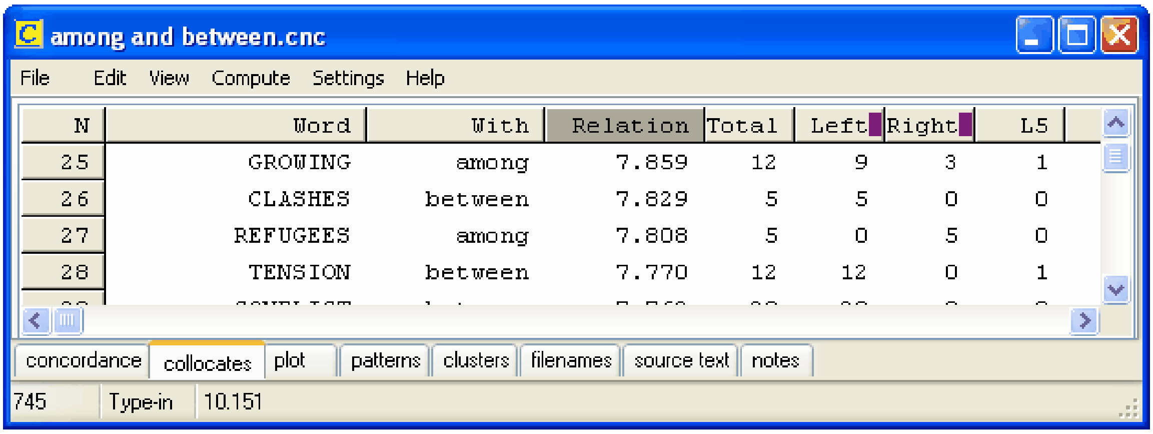
The number of columns will depend on the collocation word horizons. With 5,5 you'll get five columns to the left and 5 to the right of the search word. So you can see exactly how many times each word was found in the general neighbourhood of the search word and how many times it was found exactly 1 word to the left or 4 words to the right, for example.
The most frequent will be signalled in *most frequent collocate colour* 60 (default=**red**). In the screenshot below, **`differences`** comes 44 times in total but **39** of these are in position L1.



N	Word	With	Relation	Total	Left	Right	L5	L4	L3	L2	L1	entre	R1
1	AMONG	among	12.133	621	0	0	0	0	0	0	0	621	0
2	BETWEEN	between	10.436	2,109	14	14	7	4	2	1	0	2,081	0
3	DIFFERENTIATE	between	10.151	5	5	0	0	0	0	0	5	0	0
4	SIMILARITIES	between	9.360	13	13	0	0	0	2	0	11	0	0
5	DISTINCTIONS	between	9.092	6	6	0	1	1	1	0	3	0	0
6	DISTINCTION	between	9.008	20	20	0	0	3	1	1	15	0	0
7	TENSIONS	between	8.955	8	8	0	0	1	0	1	6	0	0
8	DIFFERENCES	between	8.943	44	43	1	1	1	0	2	39	0	0
9	DISTINGUISH	between	8.919	11	11	0	0	0	0	1	10	0	0
10	TORN	between	8.884	9	9	0	0	0	0	0	9	0	0
11	DIFFERENCE	between	8.714	96	96	0	2	1	3	3	87	0	0
12	OVERLAP	between	8.677	6	6	0	0	0	0	0	6	0	0

745 Type-in 10.151

The screenshot above shows collocation results for a concordance of **`BETWEEN/AMONG`** sorted by the *Relation* column, where items like **`differentiate, difference`** etc. are found to be most strongly related to **`between`**. Further down the listing, some links concerning **`among`** (**`growing, refugees`**) are to be seen.



among and between.cnc

N	Word	With	Relation	Total	Left	Right	L5
25	GROWING	among	7.859	12	9	3	1
26	CLASHES	between	7.829	5	5	0	0
27	REFUGEES	among	7.808	5	0	5	0
28	TENSION	between	7.770	12	12	0	1

745 Type-in 10.151

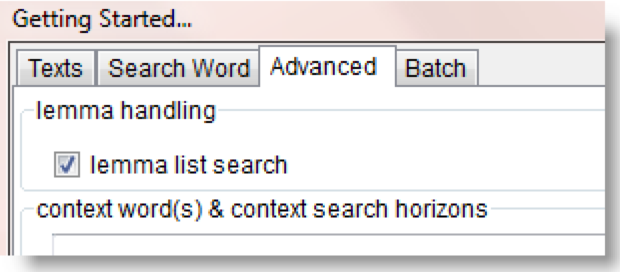
The frequency display can be re-sorted 189 ( ) and you can recalculate the collocates ( ) if you zap 129 entries from the concordance or change the horizons 180.

You can also highlight any given collocate 186 in your concordance display.

See also: Word Clouds 128, Collocation 179, Collocation Relationship 180, Collocates and Lemmas 183 , Mutual Information 289

7.9.5 collocates and lemmas

In the following case a lemma list 274 was used and lemma search specified, with a concordance on the word **abandon**:



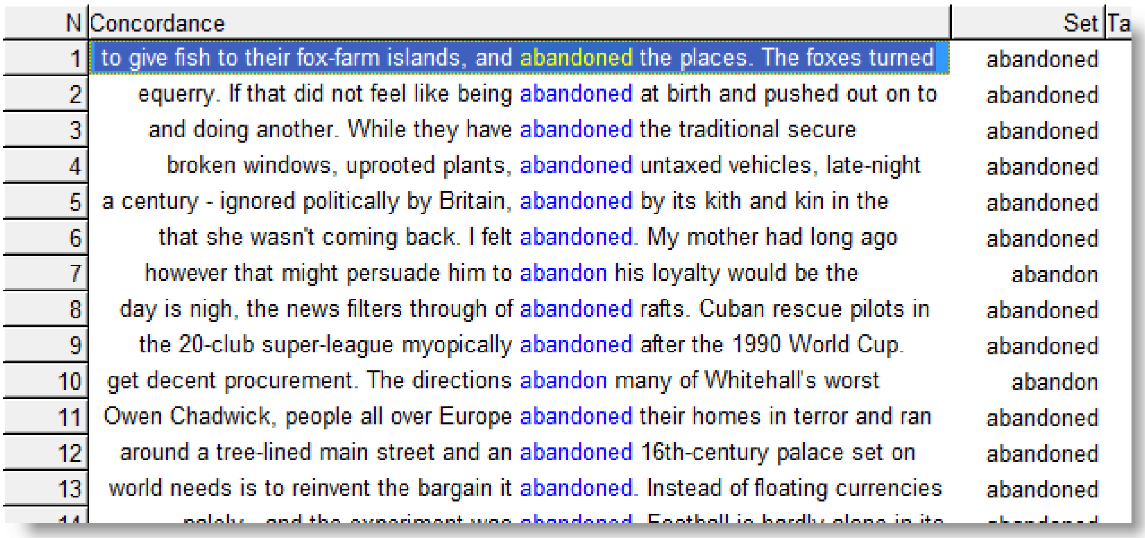
Getting Started...

lemma handling

☑ lemma list search

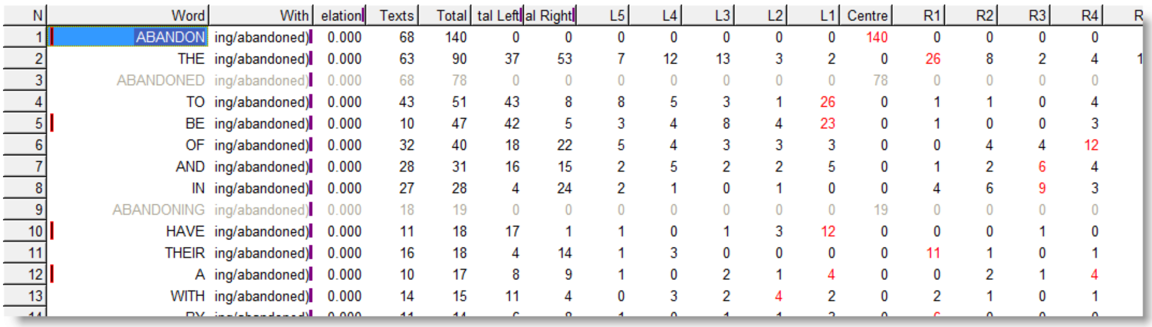
context word(s) & context search horizons

with these results showing which form of the lemma was used in the Set column.



N	Concordance	Set	Ta
1	to give fish to their fox-farm islands, and abandoned the places. The foxes turned	abandoned	
2	equerry. If that did not feel like being abandoned at birth and pushed out on to	abandoned	
3	and doing another. While they have abandoned the traditional secure	abandoned	
4	broken windows, uprooted plants, abandoned untaxed vehicles, late-night	abandoned	
5	a century - ignored politically by Britain, abandoned by its kith and kin in the	abandoned	
6	that she wasn't coming back. I felt abandoned. My mother had long ago	abandoned	
7	however that might persuade him to abandon his loyalty would be the	abandon	
8	day is nigh, the news filters through of abandoned rafts. Cuban rescue pilots in	abandoned	
9	the 20-club super-league myopically abandoned after the 1990 World Cup.	abandoned	
10	get decent procurement. The directions abandon many of Whitehall's worst	abandon	
11	Owen Chadwick, people all over Europe abandoned their homes in terror and ran	abandoned	
12	around a tree-lined main street and an abandoned 16th-century palace set on	abandoned	
13	world needs is to reinvent the bargain it abandoned. Instead of floating currencies	abandoned	

In the collocate window below, the red line in row 1 indicates that the 140 cases of **ABANDON** include other forms such as 78 cases of **ABANDONED** and 19 of **ABANDONING** (greyed out below).



N	Word	With	elation	Texts	Total	tal Left	al Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R
1	ABANDON	ing/abandoned)	0.000	68	140	0	0	0	0	0	0	0	140	0	0	0	0	
2	THE	ing/abandoned)	0.000	63	90	37	53	7	12	13	3	2	0	26	8	2	4	1
3	ABANDONED	ing/abandoned)	0.000	68	78	0	0	0	0	0	0	0	78	0	0	0	0	
4	TO	ing/abandoned)	0.000	43	51	43	8	8	5	3	1	26	0	1	1	0	4	
5	BE	ing/abandoned)	0.000	10	47	42	5	3	4	8	4	23	0	1	0	0	3	
6	OF	ing/abandoned)	0.000	32	40	18	22	5	4	3	3	3	0	0	4	4	12	
7	AND	ing/abandoned)	0.000	28	31	16	15	2	5	2	2	5	0	1	2	6	4	
8	IN	ing/abandoned)	0.000	27	28	4	24	2	1	0	1	0	0	4	6	9	3	
9	ABANDONING	ing/abandoned)	0.000	18	19	0	0	0	0	0	0	0	19	0	0	0	0	
10	HAVE	ing/abandoned)	0.000	11	18	17	1	1	0	1	3	12	0	0	0	1	0	
11	THEIR	ing/abandoned)	0.000	16	18	4	14	1	3	0	0	0	0	11	1	0	1	
12	A	ing/abandoned)	0.000	10	17	8	9	1	0	2	1	4	0	0	2	1	4	
13	WITH	ing/abandoned)	0.000	14	15	11	4	0	3	2	4	2	0	2	1	0	1	

The red mark by **BE** (row 5) shows that this row gives collocation numbers covering all forms of BE such as WAS, WERE etc. Similarly, **HAVE** and **A** are lemmatised in this screenshot.

Thus, for your search-word and its variants you can see detailed frequencies, but its collocates, though they do get lemmatised, do not show you the variant forms or any specific frequencies.

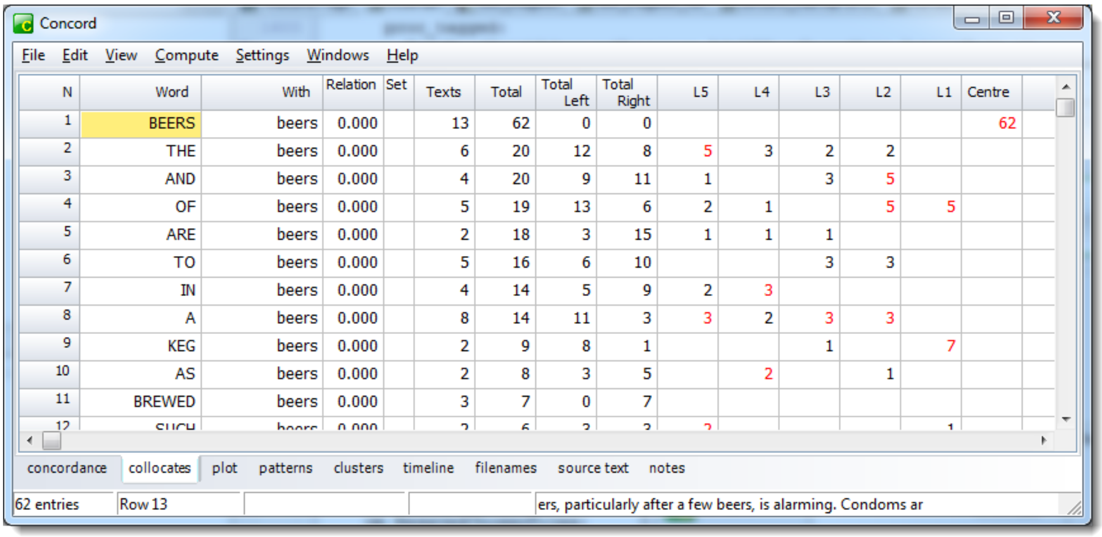
7.9.6 collocate follow

The point of it

The idea (from Paul Raper) is to be able to follow up a collocate by requesting a new concordance based on it, in the same text files as selected for the collocate. This aids exploration of related words.

How to do it

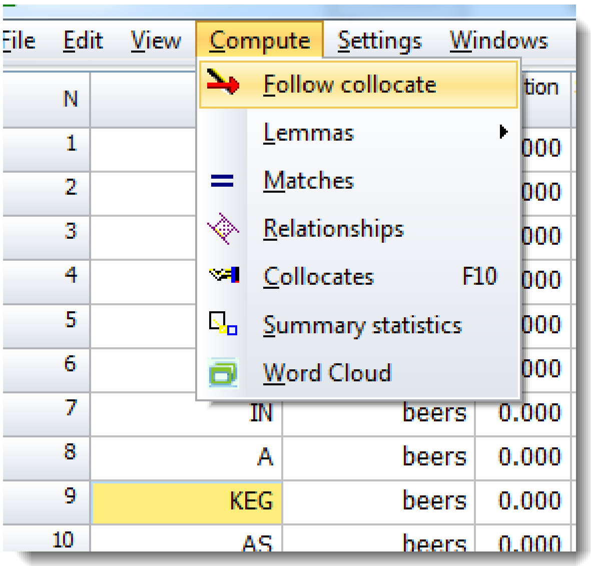
Here is an example, where there is a collocate list relating to **BEERS**. Select a word of interest, such as **KEG,**



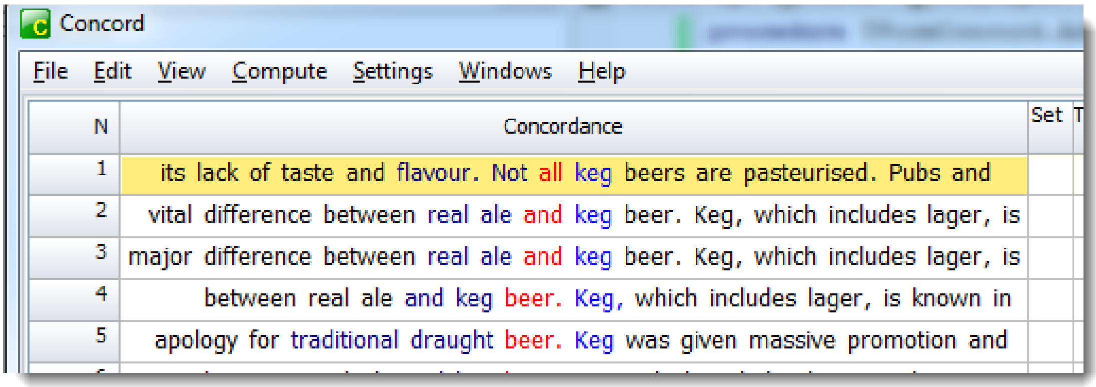
N	Word	With	Relation	Set	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre
1	BEERS	beers	0.000		13	62	0	0						62
2	THE	beers	0.000		6	20	12	8	5	3	2	2		
3	AND	beers	0.000		4	20	9	11	1		3	5		
4	OF	beers	0.000		5	19	13	6	2	1		5	5	
5	ARE	beers	0.000		2	18	3	15	1	1	1			
6	TO	beers	0.000		5	16	6	10			3	3		
7	IN	beers	0.000		4	14	5	9	2	3				
8	A	beers	0.000		8	14	11	3	3	2	3	3		
9	KEG	beers	0.000		2	9	8	1			1		7	
10	AS	beers	0.000		2	8	3	5		2		1		
11	BREWED	beers	0.000		3	7	0	7						
12	SUCH	beers	0.000		2	6	3	3	2				1	

62 entries | Row 13 | ers, particularly after a few beers, is alarming. Condoms ar

and select *Follow collocate* in the *Compute* menu.



WordSmith starts up a new Concord window with a search on KEG.



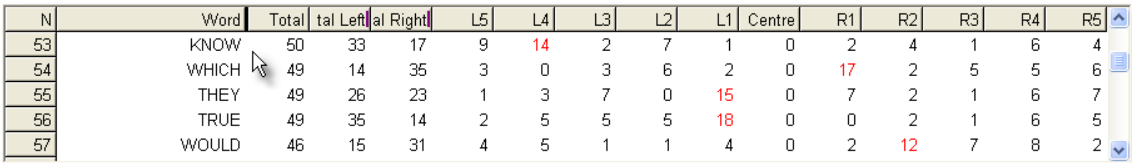
N	Concordance	Set	T
1	its lack of taste and flavour. Not all keg beers are pasteurised. Pubs and		
2	vital difference between real ale and keg beer. Keg, which includes lager, is		
3	major difference between real ale and keg beer. Keg, which includes lager, is		
4	between real ale and keg beer. Keg, which includes lager, is known in		
5	apology for traditional draught beer. Keg was given massive promotion and		

The search is carried out on the most recently selected text files (selected using the file-choose window 44) or by reading in a saved concordance).

7.9.7 collocate highlighting in concordance

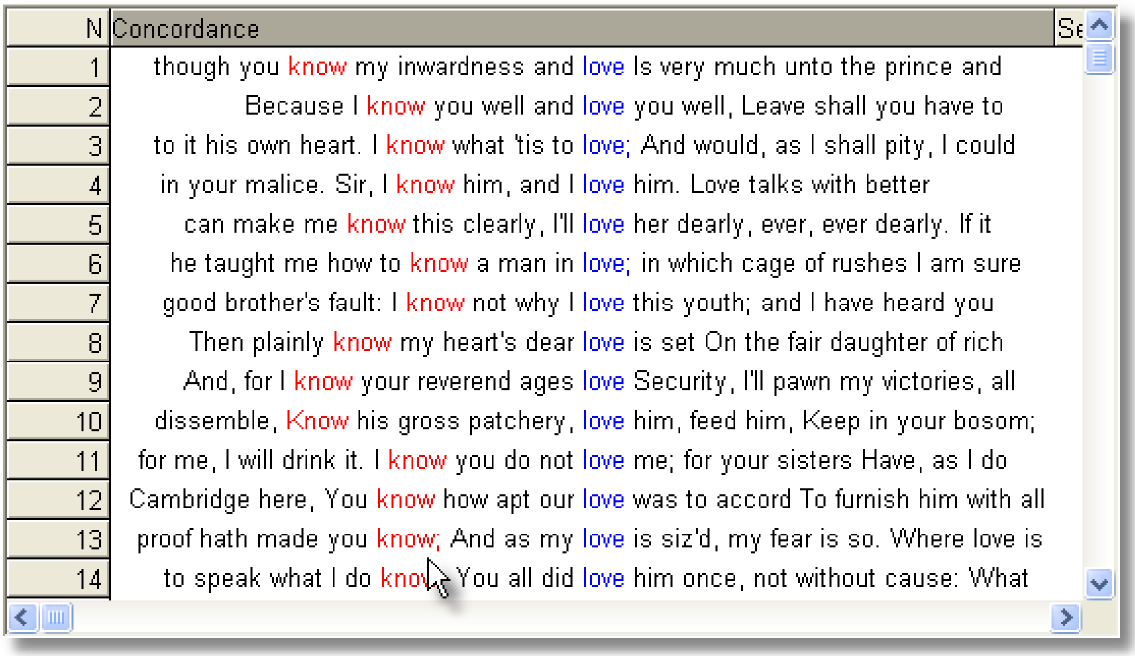
The point of it...

The idea is to be able to see a selected collocate highlighted in the concordance. In this example, the texts were Shakespeare plays and search word was **love**. One of the collocates is **know**, occurring a total of 50 times, with the most frequent at position 4 words to the left of **love**.



N	Word	Total	tal Left	al Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
53	KNOW	50	33	17	9	14	2	7	1	0	2	4	1	6	4
54	WHICH	49	14	35	3	0	3	6	2	0	17	2	5	5	6
55	THEY	49	26	23	1	3	7	0	15	0	7	2	1	6	7
56	TRUE	49	35	14	2	5	5	5	18	0	0	2	1	6	5
57	WOULD	46	15	31	4	5	1	1	4	0	2	12	7	8	2

Double-clicking **14** in the L4 column to the right of **know**, we see this in the concordance:



N	Concordance	Se
1	though you know my inwardness and love Is very much unto the prince and	
2	Because I know you well and love you well, Leave shall you have to	
3	to it his own heart. I know what 'tis to love; And would, as I shall pity, I could	
4	in your malice. Sir, I know him, and I love him. Love talks with better	
5	can make me know this clearly, I'll love her dearly, ever, ever dearly. If it	
6	he taught me how to know a man in love; in which cage of rushes I am sure	
7	good brother's fault: I know not why I love this youth; and I have heard you	
8	Then plainly know my heart's dear love is set On the fair daughter of rich	
9	And, for I know your reverend ages love Security, I'll pawn my victories, all	
10	dissemble, Know his gross patchery, love him, feed him, Keep in your bosom;	
11	for me, I will drink it. I know you do not love me; for your sisters Have, as I do	
12	Cambridge here, You know how apt our love was to accord To furnish him with all	
13	proof hath made you know; And as my love is siz'd, my fear is so. Where love is	
14	to speak what I do know You all did love him once, not without cause: What	

We have brought to the top of the concordance those lines which contain **know** in position L4.

How to do it

In a collocates window or a patterns window, simply double-click the item you wish to highlight. Or select it and choose *View | Highlight selected collocate*.

In the collocates window, if you click

	what you get
the Word column or the Total column	all instances of the word
Total Left	those to the left (33 in the case of **know** above)
Total Right	those to the right (17)
otherwise	those in that column only

To get rid

Re-sort[208] in a different way or choose the menu item *View | Refresh*.

7.9.8 collocate settings

To set collocation horizons and other **Concord** settings, in the main **WordSmith** Controller[4] menu at the top, choose *Concord Settings*.

Collocates are computed case-insensitively (so **my** in the concordance line will be treated like **My**).
If you don't want certain collocates such as **THE** to be included, use a stop-list[120].
You can lemmatise (join related forms like **SPEAK -> SPEAKS, SPOKE, SPOKEN**) using a

lemma list file 274.

Minimum Specifications

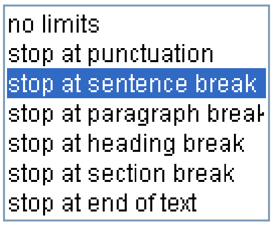
The minimum length is 1, and minimum frequency is 1 (default is 10). You can specify here how frequently it must have appeared in the neighbourhood of the Search Word. Words which only come once or twice are less likely to be informative. So specifying 5 will only show a collocate which comes 5 or more times in the neighbouring context.
Similarly, you can specify how long a collocate must be for it to be stored in memory, e.g. 3 letters or more would be 3.

Horizons

Here you specify how many words to left and right of the Search Word are to be included in the collocation search: the size of the "neighbourhood" referred to above. The maximum is 25 left and 25 right. Results will later show you these in separate columns so you can examine exactly how many times a given collocate cropped up say 3 words to the left of your Search Word.
The most frequent will be signalled in the *most frequent collocate colour* 60 (default=**red**).

Breaks

These are



no limits
stop at punctuation
stop at sentence break
stop at paragraph break
stop at heading break
stop at section break
stop at end of text

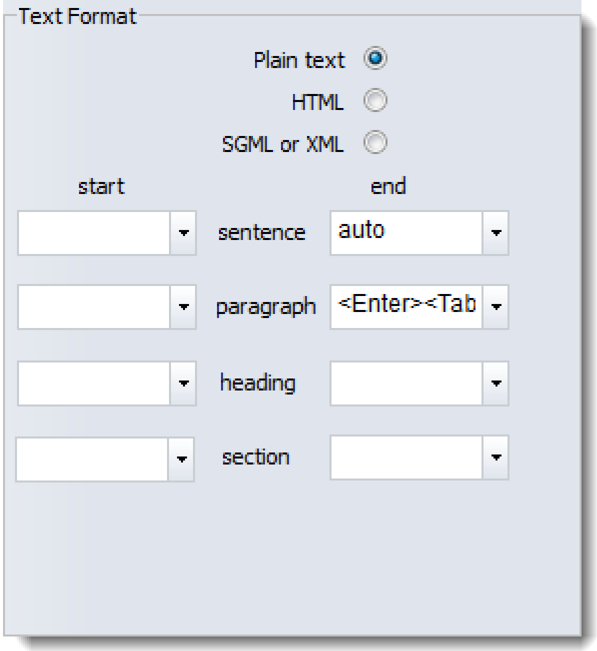
which you will see in the bottom right corner of the screen visible in the Controller Concord Settings 222.
When the collocates are computed, if the setting is to stop at sentence breaks, collocates will be counted within the above horizons but taking sentence breaks into account.

For example, if a concordance line contains

source, per pointing integration times, respectively. However, when we compared these two maps

and the search-word is **however**,
only
when we compared these two
will be used for collocates because there is a sentence break to the left of the search word. If the setting is "stop at punctuation", then nothing will come into the collocate list for that line (because there is a more major break than punctuation to the left of it, and no word to the right of the search-word before a punctuation symbol.

stop at end of text: end of text is by default assumed to be the end of the text file. *stop at heading* or *section*: this works by recognising ends of heading or section which you can specify in the text format box (language settings):



Text Format

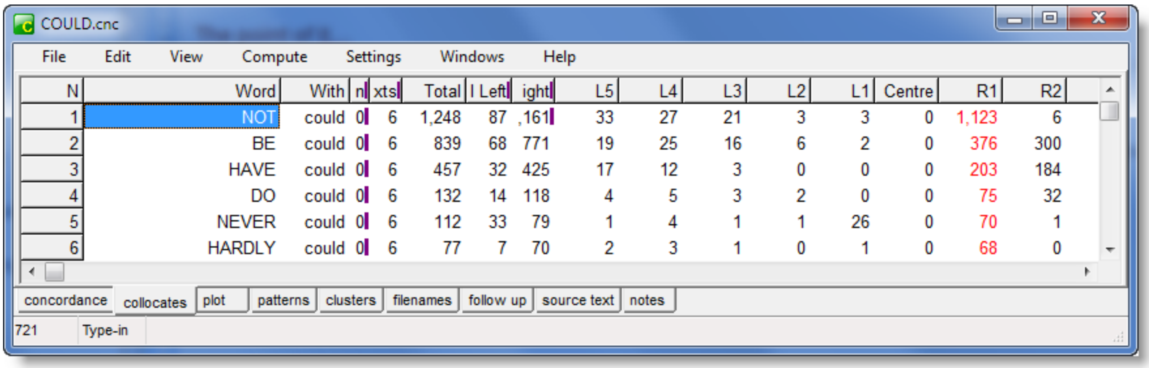
● Plain text
○ HTML
○ SGML or XML

start		end
	sentence	auto
	paragraph	<Enter><Tab
	heading	
	section	

7.9.9 re-sorting: collocates

The point of it...

is to home in, for example, on the ones in L1 or R1 position. To find sub-patterns of collocation, so as to more fully understand the company your search-word keeps.



N	Word	With	n	xts	Total	l Left	ight	L5	L4	L3	L2	L1	Centre	R1	R2
1	NOT	could	0	6	1,248	87	,161	33	27	21	3	3	0	1,123	6
2	BE	could	0	6	839	68	771	19	25	16	6	2	0	376	300
3	HAVE	could	0	6	457	32	425	17	12	3	0	0	0	203	184
4	DO	could	0	6	132	14	118	4	5	3	2	0	0	75	32
5	NEVER	could	0	6	112	33	79	1	4	1	1	26	0	70	1
6	HARDLY	could	0	6	77	7	70	2	3	1	0	1	0	68	0

Here the collocates of **COULD** in some Jane Austen texts show how negatives crop up a lot in R1 position.

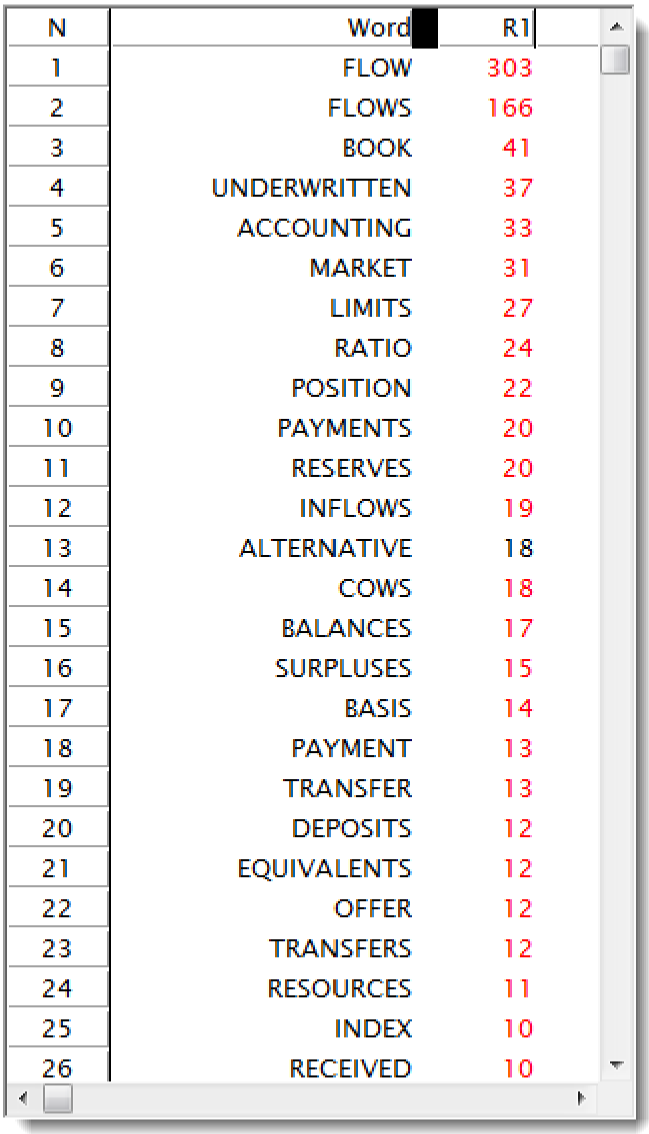
How to do it... just press the header

The frequency-ordered collocation display can be re-sorted to reveal the frequencies sorted by their total frequencies overall (the default), by the left or right frequency total, or by any individual frequency position. Just press the header of a column to sort it. Press again to toggle the sort between ascending and descending.

You can also get the concordance lines sorted so as to highlight specific collocates 186, as in the case of the 70 cases of **NEVER** in R1 position in the screenshot.

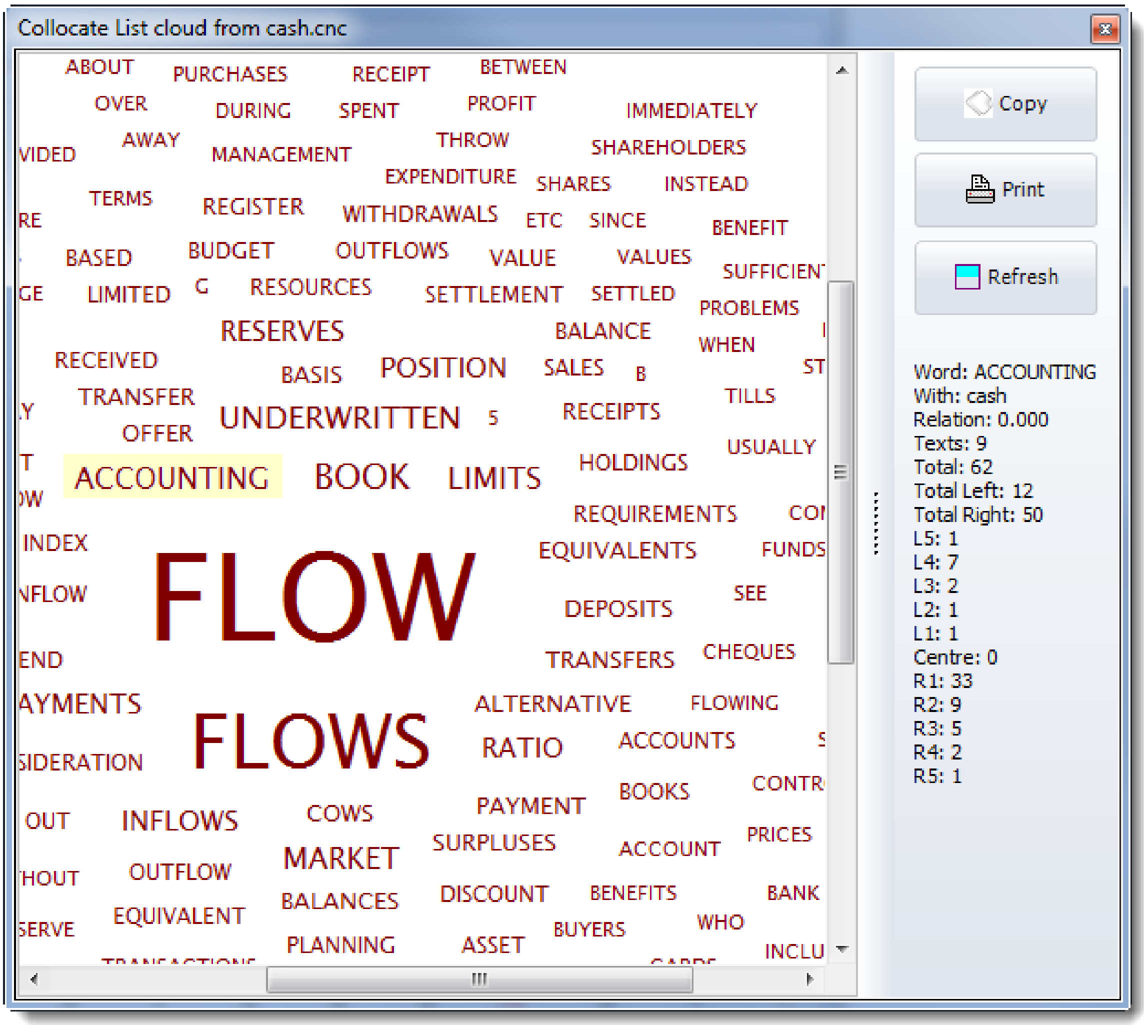
Word Clouds

You can also get a word cloud 128 of your sorted column. In the screenshot below, a concordance on **cash** generated these R1 collocates (with most function words eliminated using a stoplist 120):



N	Word	R1
1	FLOW	303
2	FLOWS	166
3	BOOK	41
4	UNDERWRITTEN	37
5	ACCOUNTING	33
6	MARKET	31
7	LIMITS	27
8	RATIO	24
9	POSITION	22
10	PAYMENTS	20
11	RESERVES	20
12	INFLOWS	19
13	ALTERNATIVE	18
14	COWS	18
15	BALANCES	17
16	SURPLUSES	15
17	BASIS	14
18	PAYMENT	13
19	TRANSFER	13
20	DEPOSITS	12
21	EQUIVALENTS	12
22	OFFER	12
23	TRANSFERS	12
24	RESOURCES	11
25	INDEX	10
26	RECEIVED	10

and these data fed straight into a word cloud.



Word: ACCOUNTING
With: cash
Relation: 0.000
Texts: 9
Total: 62
Total Left: 12
Total Right: 50
L5: 1
L4: 7
L3: 2
L2: 1
L1: 1
Centre: 0
R1: 33
R2: 9
R3: 5
R4: 2
R5: 1

In the word cloud, the mouse hovered over the word **account ing** so the details of that word are shown to the right.

See also: Collocation 179, Collocation Display 181, Collation Horizons 180, Word Clouds 128, Patterns 207.

7.10 dispersion plot

The point of it…

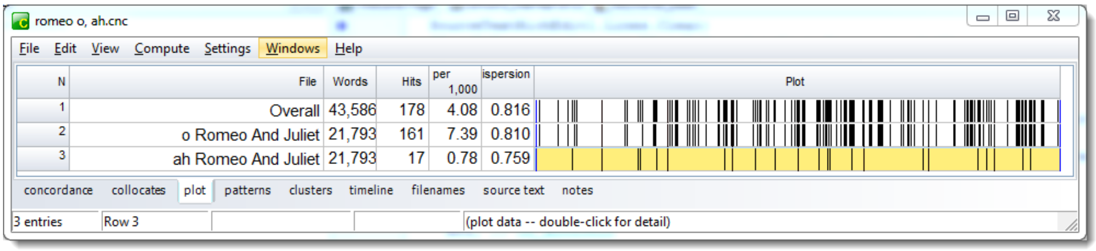
This shows where the search word occurs in the file which the current entry belongs to. That way you can see where mention is made most of your search word in each file. Another case where the aim is to promote the noticing of linguistic patterning 450.

What you see

The plot shows:

File	source text file-name
Words	number of words in the source text
Hits	number of occurrences of the search-word

per 1,000 how many occurrences per 1,000 words
Dispersion the plot dispersion value 446
Plot a plot showing where they cropped up, where the left edge of the plot represents the beginning of the text file ("Once upon a time" for example) and the right edge is at the end ("happily ever after". Though not in the case of Romeo and Juliet.).

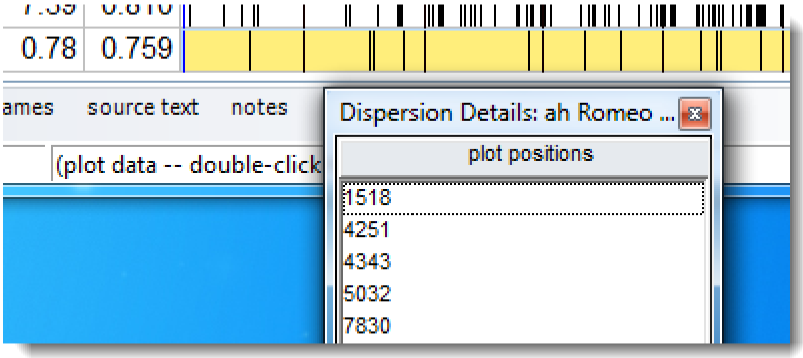

romeo o, ah.cnc
File Edit View Compute Settings Windows Help
concordance collocates plot patterns clusters timeline filenames source text notes

N	File	Words	Hits	per 1,000	ispersion	Plot
1	Overall	43,586	178	4.08	0.816	
2	o Romeo And Juliet	21,793	161	7.39	0.810	
3	ah Romeo And Juliet	21,793	17	0.78	0.759	

3 entries | Row 3 | (plot data -- double-click for detail)

Here we see a plot of "O" and another of "AH" from the play Romeo and Juliet. They are on separate lines because there were 2 search-words. There are more "O" exclamations than "AH"s.

As the status bar says, you can get the word numbers for the plot by double-clicking the plot area:



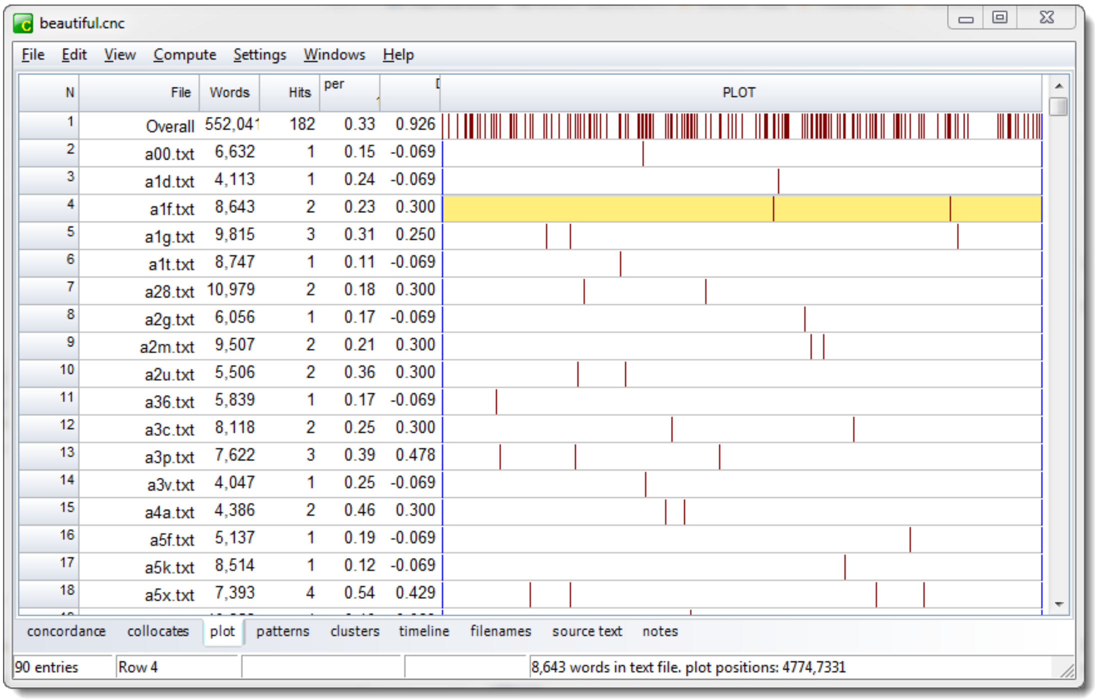
0.78 0.759

plot positions

1518
4251
4343
5032
7830

Using *View* | *Ruler*, you can switch on a "ruler" splitting the display into segments.

The plot below is of one search-word (`beautiful`) in lots of texts.


beautiful.cnc
File Edit View Compute Settings Windows Help
concordance collocates plot patterns clusters timeline filenames source text notes

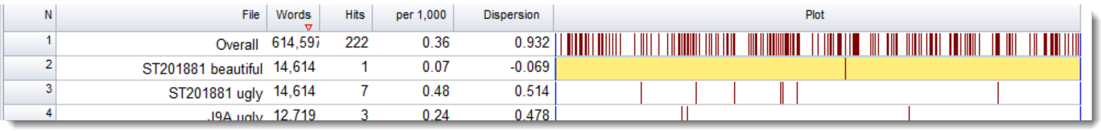
N	File	Words	Hits	per	D	PLOT
1	Overall	552,04	182	0.33	0.926	
2	a00.txt	6,632	1	0.15	-0.069	
3	a1d.txt	4,113	1	0.24	-0.069	
4	a1f.txt	8,643	2	0.23	0.300	
5	a1g.txt	9,815	3	0.31	0.250	
6	a1t.txt	8,747	1	0.11	-0.069	
7	a28.txt	10,979	2	0.18	0.300	
8	a2g.txt	6,056	1	0.17	-0.069	
9	a2m.txt	9,507	2	0.21	0.300	
10	a2u.txt	5,506	2	0.36	0.300	
11	a36.txt	5,839	1	0.17	-0.069	
12	a3c.txt	8,118	2	0.25	0.300	
13	a3p.txt	7,622	3	0.39	0.478	
14	a3v.txt	4,047	1	0.25	-0.069	
15	a4a.txt	4,386	2	0.46	0.300	
16	a5f.txt	5,137	1	0.19	-0.069	
17	a5k.txt	8,514	1	0.12	-0.069	
18	a5x.txt	7,393	4	0.54	0.429	

90 entries | Row 4 | 8,643 words in text file. plot positions: 4774,7331

The status-bar gives details of the highlighted text.

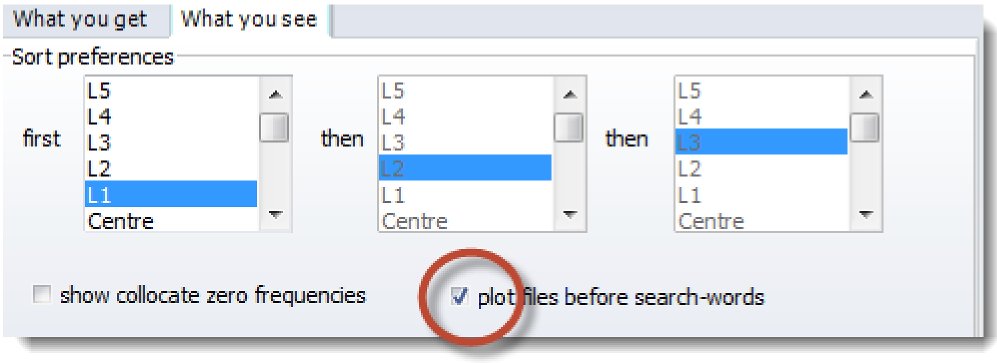
Multiple Search-words or Texts

If there are 2 or more search-words or texts, you will see something like this:



N	File	Words	Hits	per 1,000	Dispersion	Plot
1	Overall	614,597	222	0.36	0.932	
2	ST201881 beautiful	14,614	1	0.07	-0.069	
3	ST201881 ugly	14,614	7	0.48	0.514	
4	J9A ugly	12,719	3	0.24	0.478	

where the *File* column supplies the file-name and the search-word in that order. If you want it with the search-word first, go to the Concord settings in the Controller, What you see, and click here:



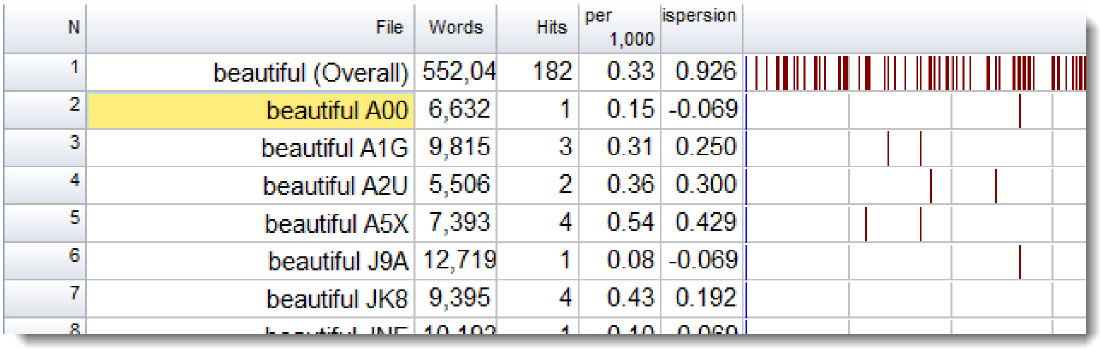
Sort preferences

first L1 then L2 then L3

□ show collocate zero frequencies

☑ plot files before search-words

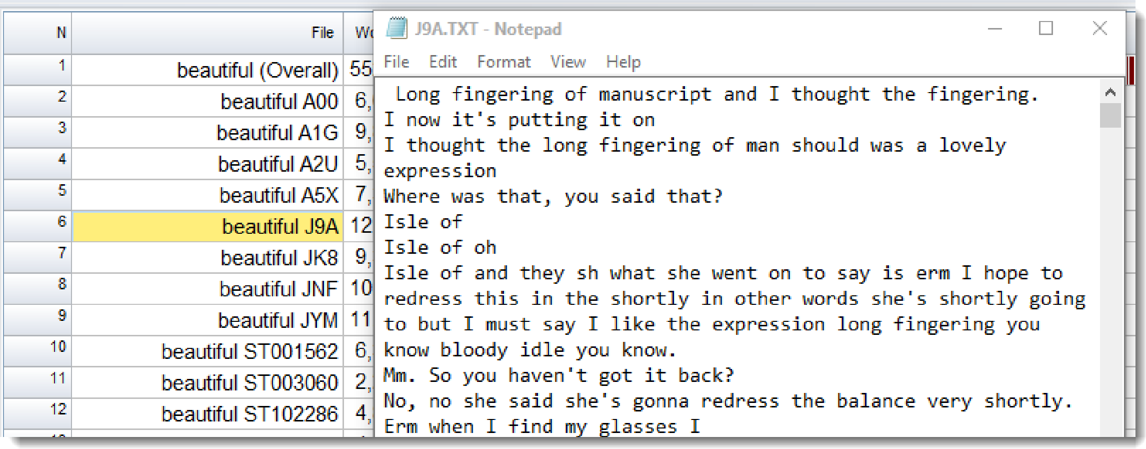
and re-sort the File list:



N	File	Words	Hits	per 1,000	ispersion
1	beautiful (Overall)	552,04	182	0.33	0.926
2	beautiful A00	6,632	1	0.15	-0.069
3	beautiful A1G	9,815	3	0.31	0.250
4	beautiful A2U	5,506	2	0.36	0.300
5	beautiful A5X	7,393	4	0.54	0.429
6	beautiful J9A	12,719	1	0.08	-0.069
7	beautiful JK8	9,395	4	0.43	0.192
8	beautiful JNF	10,102	1	0.10	0.069

Double-click to see the source text

Just double-click in the File column:

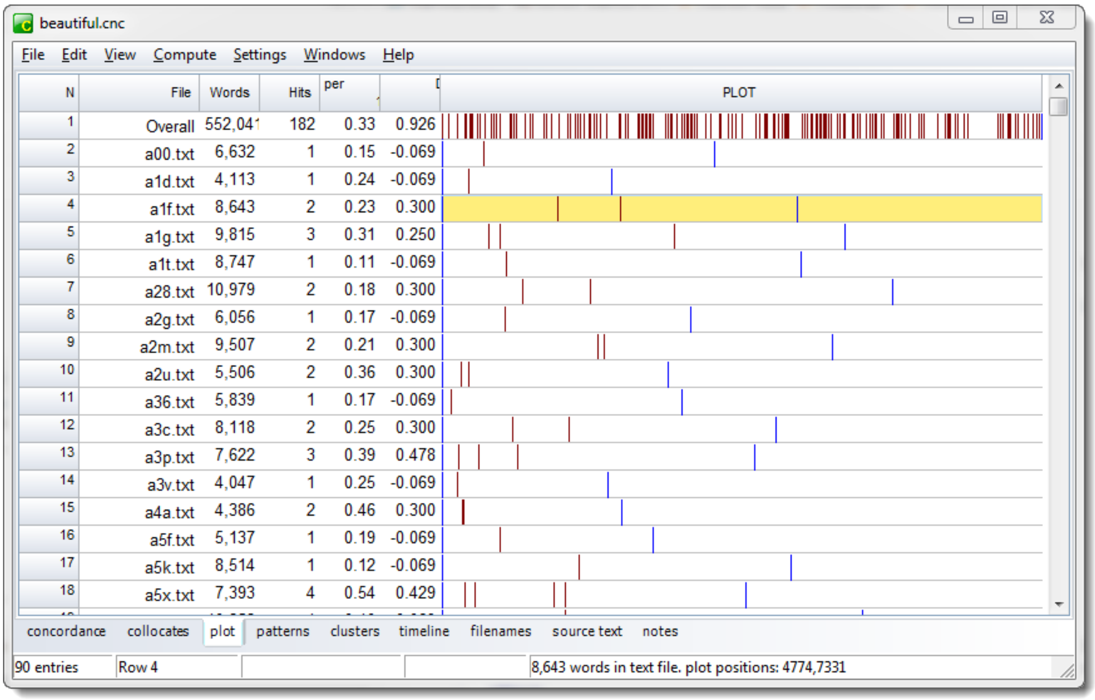


N	File	Wo
1	beautiful (Overall)	55
2	beautiful A00	6,
3	beautiful A1G	9,
4	beautiful A2U	5,
5	beautiful A5X	7,
6	beautiful J9A	12
7	beautiful JK8	9,
8	beautiful JNF	10
9	beautiful JYM	11
10	beautiful ST001562	6,
11	beautiful ST003060	2,
12	beautiful ST102286	4,

 Long fingering of manuscript and I thought the fingering.
I now it's putting it on
I thought the long fingering of man should was a lovely
expression
Where was that, you said that?
Isle of
Isle of oh
Isle of and they sh what she went on to say is erm I hope to
redress this in the shortly in other words she's shortly going
to but I must say I like the expression long fingering you
know bloody idle you know.
Mm. So you haven't got it back?
No, no she said she's gonna redress the balance very shortly.
Erm when I find my glasses I

Uniform view

There are two ways of viewing the plot, the default, where all plotting rectangles are the same length, or Uniform Plot (where the plot rectangles reflect the original file size -- the biggest file is longest). Change this in the View menu at the top. Here is the same one with Uniform plot. The blue edge at the right reflects the file size in each case.


beautiful.cnc
File Edit View Compute Settings Windows Help
concordance collocates plot patterns clusters timeline filenames source text notes

N	File	Words	Hits	per	[	PLOT
1	Overall	552,041	182	0.33	0.926	
2	a00.txt	6,632	1	0.15	-0.069	
3	a1d.txt	4,113	1	0.24	-0.069	
4	a1f.txt	8,643	2	0.23	0.300	
5	a1g.txt	9,815	3	0.31	0.250	
6	a1t.txt	8,747	1	0.11	-0.069	
7	a28.txt	10,979	2	0.18	0.300	
8	a2g.txt	6,056	1	0.17	-0.069	
9	a2m.txt	9,507	2	0.21	0.300	
10	a2u.txt	5,506	2	0.36	0.300	
11	a36.txt	5,839	1	0.17	-0.069	
12	a3c.txt	8,118	2	0.25	0.300	
13	a3p.txt	7,622	3	0.39	0.478	
14	a3v.txt	4,047	1	0.25	-0.069	
15	a4a.txt	4,386	2	0.46	0.300	
16	a5f.txt	5,137	1	0.19	-0.069	
17	a5k.txt	8,514	1	0.12	-0.069	
18	a5x.txt	7,393	4	0.54	0.429	

90 entries | Row 4 | 8,643 words in text file. plot positions: 4774,7331

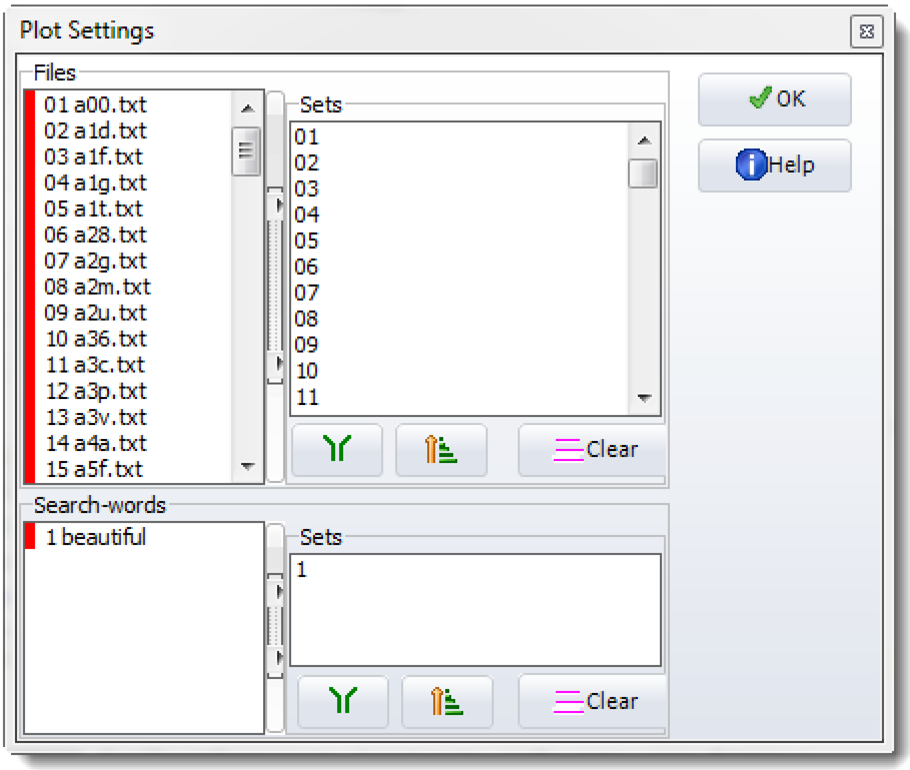
If you don't see as many marks as the number of hits, that'll be because the hits came too close together for the amount of screen space in proportion to your screen resolution. You can stretch the plot by dragging the top right edge of it. You can export the plot using Save As 102 and can get your spreadsheet to make graphs etc, as explained here 102.

Each plot window is dependent on the concordance from which it was derived. If you close the original concordance down, it will disappear. You can *Print* the plot. There's no *Save* option for the plot alone but you can of course save the concordance itself. You can *Copy* to the clipboard 422 (Ctrl+C) and then put it into a word processor as a graphic, using Paste Special.

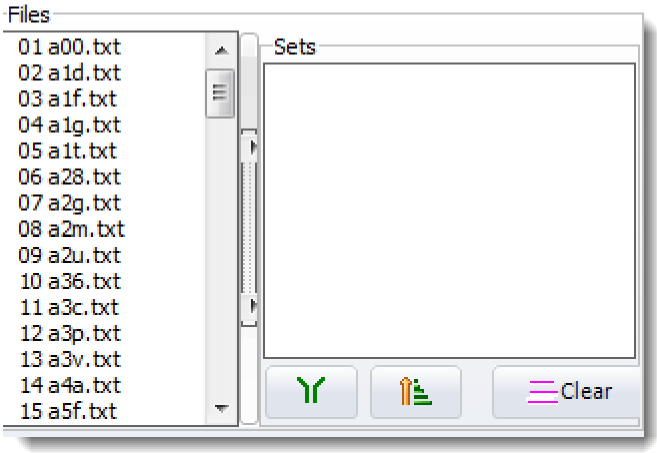
Advanced plots

When you first compute a concordance, the plot will assume you want a dispersion plot of each text file on a separate line and each different search-word on a separate line as seen above. If you have more than one text file or search-word, when you choose the *Compute* | *Plot* menu item afterwards, you will get a chance to merge your plots and omit some text files or search-words.

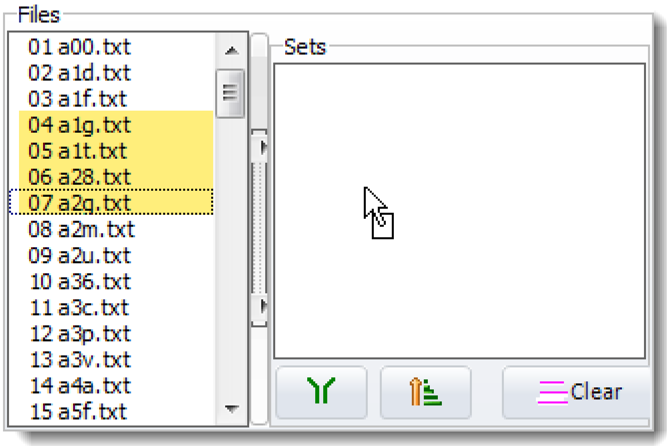
A first view of the plot settings may resemble this. All the files have by default been sorted into separate sets and so have all the search-words. The red colour indicates files or search-words which have been included in each list of sets at the right.



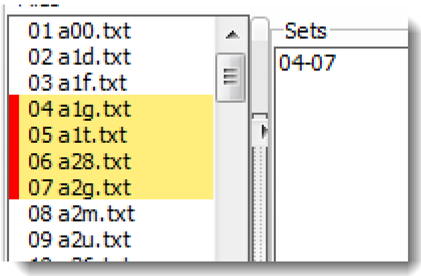
Now if you *Clear* them,



you can either select and drag or select and press the central button to get your preferred selections. (The button showing a green funnel will put all into one set, the other one will use one set for each, by the way.)



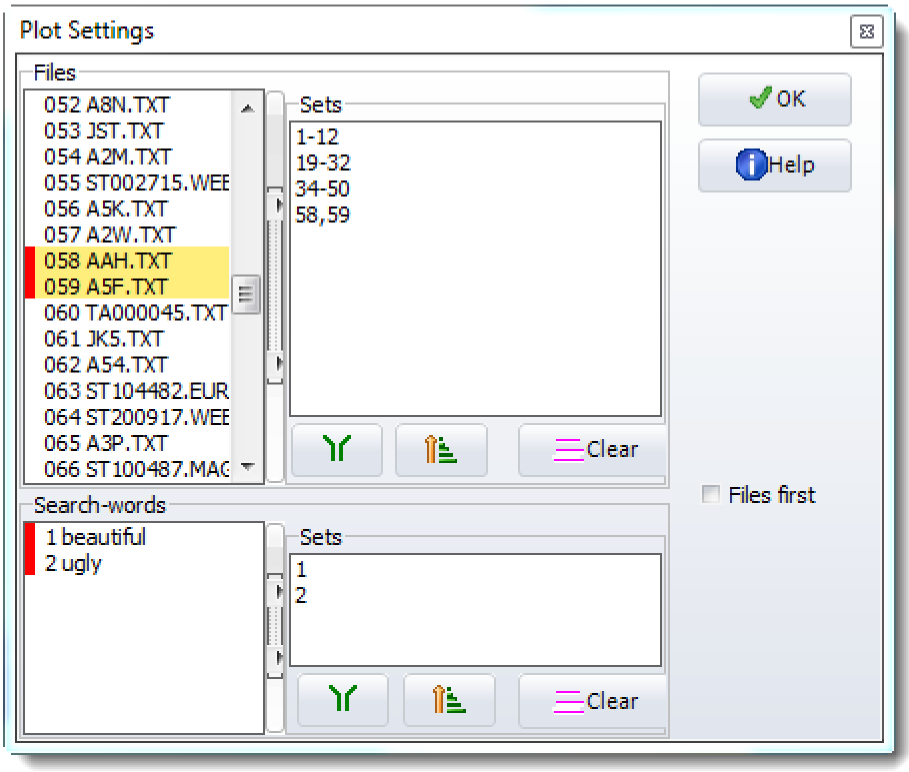
01 a00.txt
02 a1d.txt
03 a1f.txt
04 a1g.txt
05 a1t.txt
06 a28.txt
07 a2g.txt
08 a2m.txt
09 a2u.txt
10 a36.txt
11 a3c.txt
12 a3p.txt
13 a3v.txt
14 a4a.txt
15 a5f.txt



01 a00.txt
02 a1d.txt
03 a1f.txt
04 a1g.txt
05 a1t.txt
06 a28.txt
07 a2g.txt
08 a2m.txt
09 a2u.txt

04-07

Here is a set of preferences with lots of files and two search-words:



052 A8N.TXT
053 JST.TXT
054 A2M.TXT
055 ST002715.WEE
056 A5K.TXT
057 A2W.TXT
058 AAH.TXT
059 A5F.TXT
060 TA000045.TXT
061 JK5.TXT
062 A54.TXT
063 ST104482.EUR
064 ST200917.WEE
065 A3P.TXT
066 ST100487.MAG

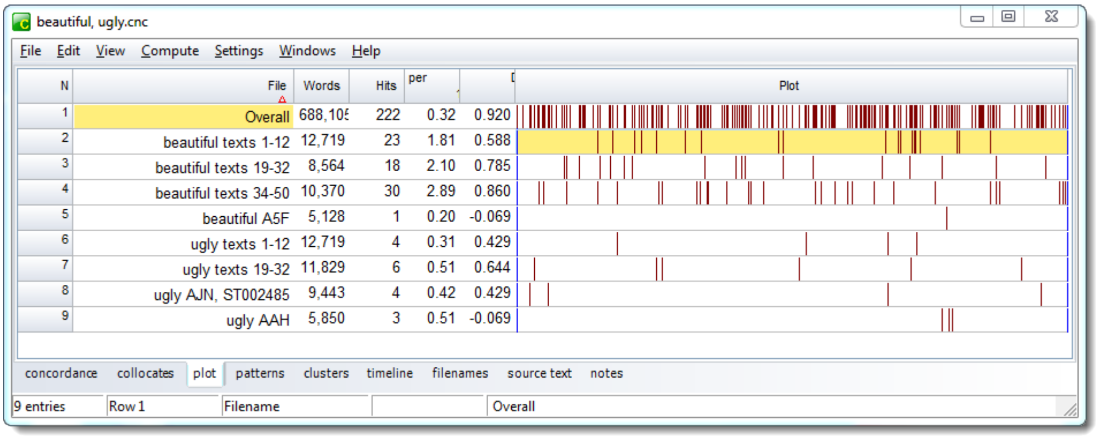
1-12
19-32
34-50
58,59

☐ Files first

1 beautiful
2 ugly

1
2

giving results like this:


beautiful, ugly.cnc
File Edit View Compute Settings Windows Help
concordance collocates plot patterns clusters timeline filenames source text notes

N	File	Words	Hits	per	[	Plot
1	Overall	688,105	222	0.32	0.920	
2	beautiful texts 1-12	12,719	23	1.81	0.588	
3	beautiful texts 19-32	8,564	18	2.10	0.785	
4	beautiful texts 34-50	10,370	30	2.89	0.860	
5	beautiful A5F	5,128	1	0.20	-0.069	
6	ugly texts 1-12	12,719	4	0.31	0.429	
7	ugly texts 19-32	11,829	6	0.51	0.644	
8	ugly AJN, ST002485	9,443	4	0.42	0.429	
9	ugly AAH	5,850	3	0.51	-0.069	

9 entries | Row 1 | Filename | | Overall

See also: plot and ruler colours 60, plot dispersion value 446.

7.11 concordancing on tags

The point of it...

Suppose you're interested in identifying structures of a certain type (as opposed to a given word or phrase), for example sequences of **Noun+Noun+Noun**. You can type in the tags you want to concordance on (with or without any words).

How to do it...

In Concord's search-word box, type in the tags you are interested in. Or define your tags in a tag-file 141.

Examples

<w NN1>table finds *table* as a singular noun (as opposed to as a verb)

<w NN1>* <w NN1>* will find any sequence of two singular common nouns in the BNC Sampler.

Note that **<w NN1>table** finds *table* if your text is tagged with **<** and **>** symbols, or if you have specified **[** and **]** as tag symbols, it will find **[w NN1]table**.

There are some more examples under Search word or phrase 159.

It doesn't matter whether you are using a tag file 141 or not, since WordSmith will identify your tags automatically. (But not by magic: of course you do need to use previously tagged text to use this function.)

In example 2, the asterisks are because in the BNC, the tags come immediately before the word they refer to: if you forgot the asterisk, Concord would assume you wanted a tag with a separator 427 on either side.

Are you concordancing on tags?

If you are asked this and your search-word or phrase includes tags, answer "Yes" to this question. If not, your search word will get " " inserted around each < or > symbol in it, as explained under

Search Word Syntax 159.

Case Sensitivity

Tags are only case sensitive if your search-word or phrase is. Search words aren't (by default). So in example 1, you will retrieve *table* and *Table* and *TABLE* if used as nouns (but nothing at all if no tags are in your source texts).

Hide Tags?

After you have generated a concordance you may wish to hide 201 the mark-up. See the View menu for this.

See also: Overview of Tags 131, Handling Tags 132, Showing Nearest Tags in Concord 199, Search word or phrase 159, Types of Tag 145, Viewing the Tags 390, Using Tags as Text Selectors 134

7.11.1 nearest tag

Concord allows you to see the nearest tag, if you have specified a tag file 141, which teaches WordSmith Tools what your preferred tags are. Then, with a concordance on screen, you'll see the tag in one of the columns of the concordance window.

The point of it...

The advantage is that you can see how your concordance search-word relates to marked-up text. For example, if you've tagged all the speech by Robert as **`[Rob]`** and Mary as **`[Mary]`**, you can quickly see in any concordance involving conversation between Mary, Robert and others, which ones came from each of them.
Alternatively, you might mark up your text as **`<Introduction>, <Body>`** and **`<Conclusion>`**: Nearest Tag will show each line like this:

```
 1 ... could not give me the time ...   <Introduction>
 2 ... Rosemary, give me another ...            <Body>
 3 ... wanted to give her the help ...          <Body>
 4 ... would not give much for that ...   <Conclusion>
```

To mark up text like this, make up a tag file 141 with your sections and label them as sections, as in these examples:

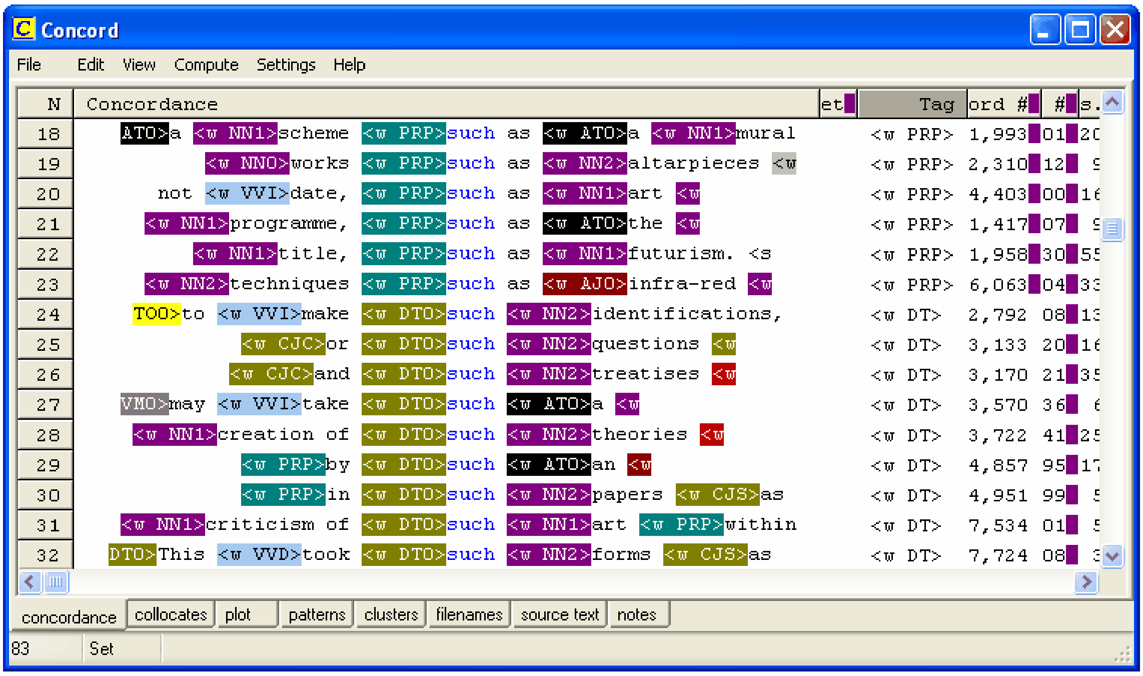
```
<ABSTRACT> /description "section"
</ABSTRACT>
<INTRODUCTION> /description "section"
</INTRODUCTION>
<SECTION 1> /description "section"
</SECTION 1>
```

or, if you want to identify the speech of all characters in a play, and have a list of the characters, and they are marked up appropriately in the text file, something like this:

```
<Romeo> /description "section"
</Romeo>
<Mercutio> /description "section"
</Mercutio>
```

<Benvolio> /description "section"
</Benvolio>
In cases using "section", Nearest Tag will find the section, however remote in the text file it may be. Without the keyword "section", Nearest Tag shows only the current context within the span of text saved 225 with each concordance line.

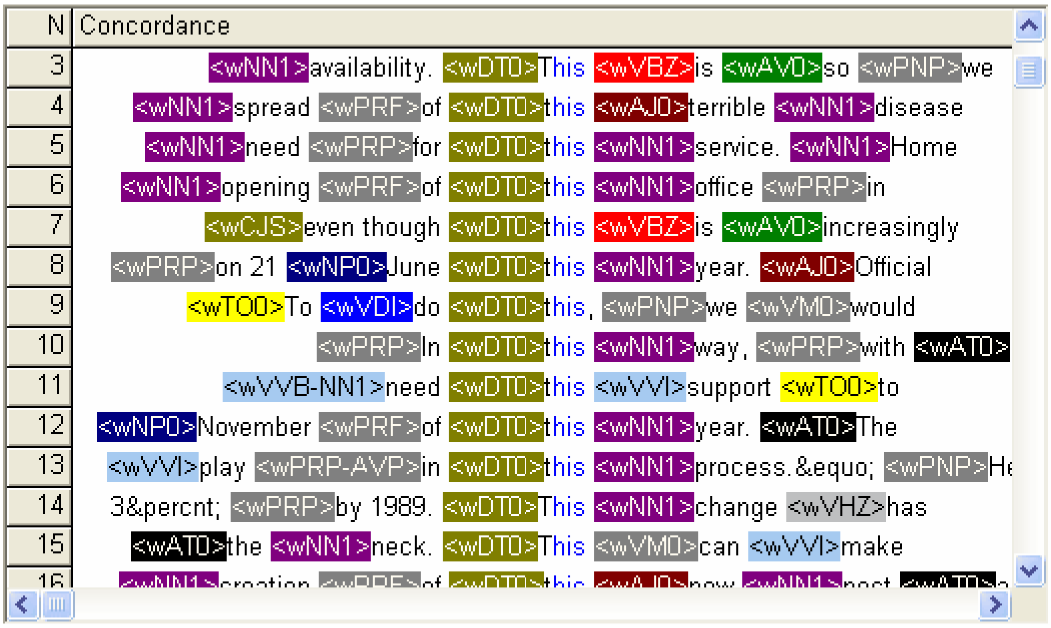
You can sort 208 on the nearest tags. In the shot below, a concordance of **such** has been computed using BNC text. Some of the cases of **such** are tagged < PRP> (**such as**) and others are <w DT0>. The Tag column shows the nearest tag, and the whole list has been sorted using that column.



N	Concordance	et	Tag	ord #	#	s.
18	ATO>a <w NN1>scheme <w PRP>such as <w ATO>a <w NN1>mural		<w PRP>	1,993	01	20
19	<w NNO>works <w PRP>such as <w NN2>altarpieces <w		<w PRP>	2,310	12	9
20	not <w VVI>date, <w PRP>such as <w NN1>art <w		<w PRP>	4,403	00	16
21	<w NN1>programme, <w PRP>such as <w ATO>the <w		<w PRP>	1,417	07	9
22	<w NN1>title, <w PRP>such as <w NN1>futurism. <s		<w PRP>	1,958	30	55
23	<w NN2>techniques <w PRP>such as <w AJO>infra-red <w		<w PRP>	6,063	04	33
24	TOO>to <w VVI>make <w DTO>such <w NN2>identifications,		<w DT>	2,792	08	13
25	<w CJC>or <w DTO>such <w NN2>questions <w		<w DT>	3,133	20	16
26	<w CJC>and <w DTO>such <w NN2>treatises <w		<w DT>	3,170	21	35
27	VMO>may <w VVI>take <w DTO>such <w ATO>a <w		<w DT>	3,570	36	6
28	<w NN1>creation of <w DTO>such <w NN2>theories <w		<w DT>	3,722	41	25
29	<w PRP>by <w DTO>such <w ATO>an <w		<w DT>	4,857	95	17
30	<w PRP>in <w DTO>such <w NN2>papers <w CJS>as		<w DT>	4,951	99	5
31	<w NN1>criticism of <w DTO>such <w NN1>art <w PRP>within		<w DT>	7,534	01	5
32	DTO>This <w VVD>took <w DTO>such <w NN2>forms <w CJS>as		<w DT>	7,724	08	3

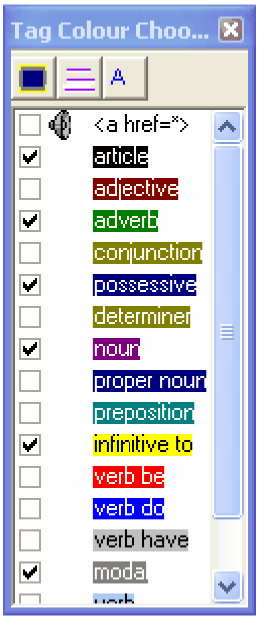
If you can't see any tags using this procedure, it is probably because the Tags to Ignore 132 have the same format. For example, if Tags to Ignore has <*>, any tags such as <title>, <quote>, etc. will be cut out of the concordance unless you specify them in a tag file 141. If so, specify the tag file and run the concordance again.

You can also display tags in colour, or even hide the tags -- yet still colour the tagged word. Here is a concordance of **this** in the BNC text with the tags in colour:



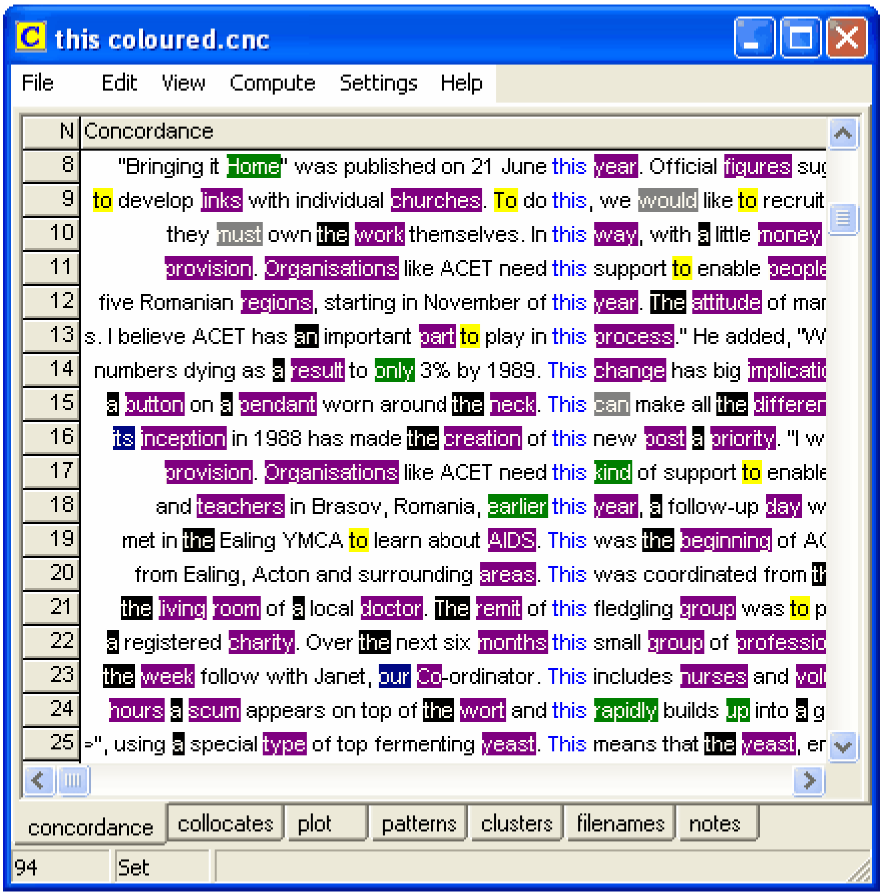
N	Concordance
3	<wNN1>availability. <wDT0>This <wVBZ>is <wAV0>so <wPNP>we
4	<wNN1>spread <wPRF>of <wDT0>this <wAJ0>terrible <wNN1>disease
5	<wNN1>need <wPRP>for <wDT0>this <wNN1>service. <wNN1>Home
6	<wNN1>opening <wPRF>of <wDT0>this <wNN1>office <wPRP>in
7	<wCJS>even though <wDT0>this <wVBZ>is <wAV0>increasingly
8	<wPRP>on 21 <wNP0>June <wDT0>this <wNN1>year. <wAJ0>Official
9	<wTO0>To <wVDI>do <wDT0>this, <wPNP>we <wVM0>would
10	<wPRP>In <wDT0>this <wNN1>way, <wPRP>with <wAT0>
11	<wVVB-NN1>need <wDT0>this <wVVI>support <wTO0>to
12	<wNP0>November <wPRF>of <wDT0>this <wNN1>year. <wAT0>The
13	<wVVI>play <wPRP-AVP>in <wDT0>this <wNN1>process.&equo; <wPNP>He
14	3&percnt; <wPRP>by 1989. <wDT0>This <wNN1>change <wVHZ>has
15	<wAT0>the <wNN1>neck. <wDT0>This <wVM0>can <wVVI>make
16	<wNN1>creation <wPRF>of <wDT0>this <wAJ0>new <wNN1>pest <wAT0>a

and here is a view showing the same data, with *View* | *Hide Tags* selected.



Tag Colour Choo...

☐ <a href=*>
☑ article
☐ adjective
☑ adverb
☐ conjunction
☑ possessive
☐ determiner
☑ noun
☐ proper noun
☐ preposition
☑ infinitive to
☐ verb be
☐ verb do
☐ verb have
☑ modal
☐ verb


this coloured.cnc
File Edit View Compute Settings Help
concordance collocates plot patterns clusters filenames notes
94 Set

N	Concordance
8	"Bringing it Home" was published on 21 June this year. Official figures sug
9	to develop links with individual churches. To do this, we would like to recruit
10	they must own the work themselves. In this way, with a little money
11	provision. Organisations like ACET need this support to enable people
12	five Romanian regions, starting in November of this year. The attitude of mar
13	s. I believe ACET has an important part to play in this process." He added, "W
14	numbers dying as a result to only 3% by 1989. This change has big implicati
15	a button on a pendant worn around the neck. This can make all the differer
16	its inception in 1988 has made the creation of this new post a priority. "I w
17	provision. Organisations like ACET need this kind of support to enable
18	and teachers in Brasov, Romania, earlier this year, a follow-up day w
19	met in the Ealing YMCA to learn about AIDS. This was the beginning of AC
20	from Ealing, Acton and surrounding areas. This was coordinated from th
21	the living room of a local doctor. The remit of this fledgling group was to p
22	a registered charity. Over the next six months this small group of professio
23	the week follow with Janet, our Co-ordinator. This includes nurses and vol
24	hours a scum appears on top of the wort and this rapidly builds up into a g
25	=", using a special type of top fermenting yeast. This means that the yeast, er

The tags themselves are no longer visible, and only 6 types of tag have been chosen to be viewed in colour.

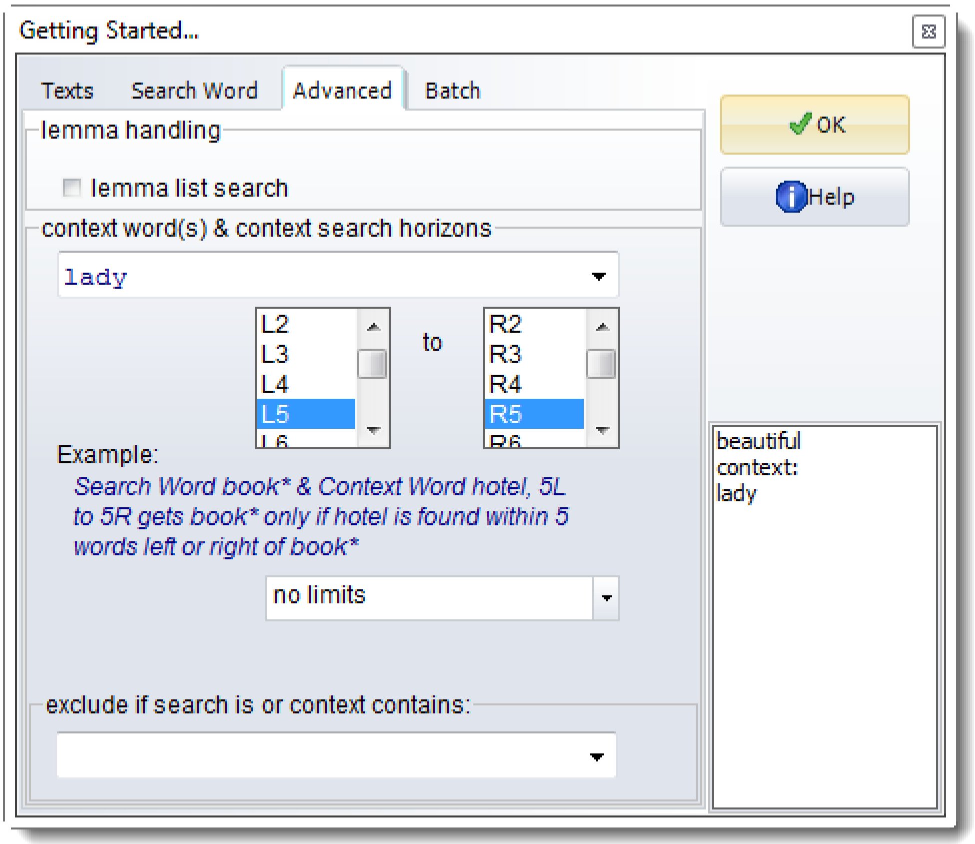
See also: Guide to handling the BNC, Overview of Tags 131, Handling Tags 132, Making a Tag File 141, Tagged Texts 131, Types of Tag 145, Viewing the Tags 390, Using Tags as Text Selectors 134

7.12 context word

You may restrict a concordance search by specifying a context word which either must or may not be present within a certain number of words of your search word.
For example, you might have **book** as your search word and **hotel*** as the context word. This will only find **book** if **hotel** or **hotels** is nearby.
Or you might have **book** as your search word and **paper*** as an exclusion criterion. This will only find **book** if **paper** or **papers** is *not* within your Context Search Horizons.

Context Search Horizons

The context horizons determine how far Concord must look to left and right of the search word when checking whether the search criteria have been met. The default 113 is 5,5 (5 to left and 5 to right of the search word) but this can be set to up to 25 on either side. 0,2 would look only to the right within two words of the search word.



□ lemma list search

context word(s) & context search horizons: lady

L5 to R5

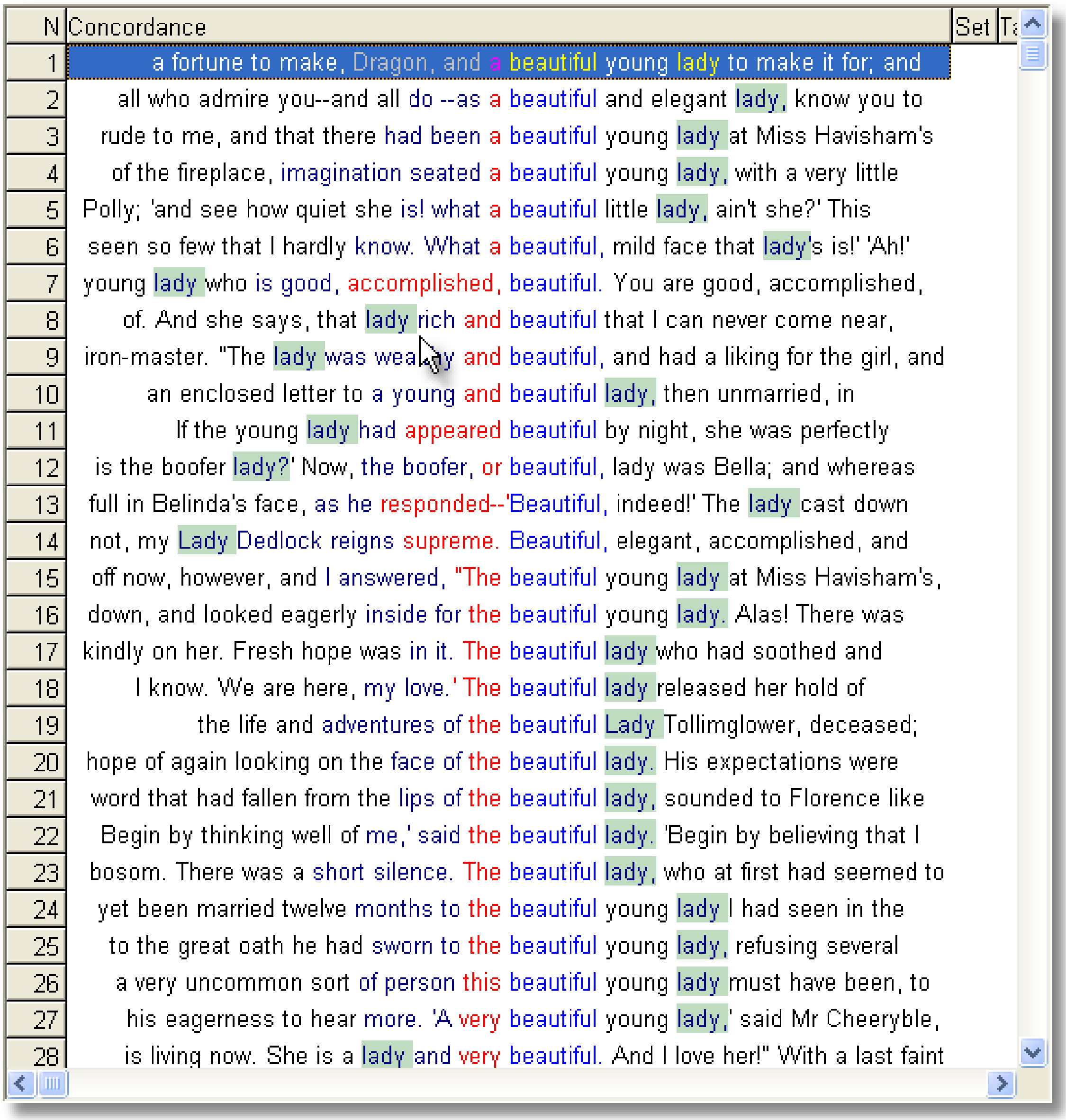
Example:
Search Word book & Context Word hotel, 5L to 5R gets book* only if hotel is found within 5 words left or right of book**

no limits

beautiful
context:
lady

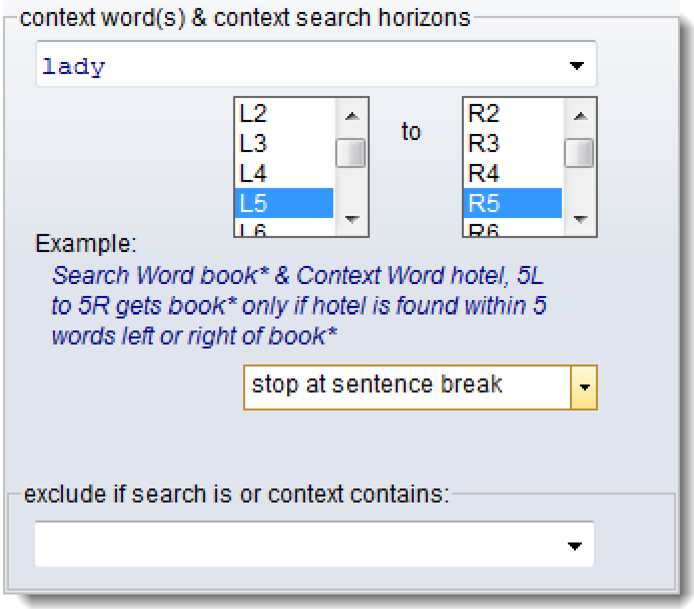
In this example the search-word is **beautiful** and the context word is **lady**, to be sought either left or right of **beautiful**.

Syntax is like that of the search word or phrase 159,
*   means disregard the end of the word and can be placed at either end of your context word.
==   means case sensitive
/   separates alternatives. You can specify up to 15 alternatives within an 80-character overall limit.
If you want to use *, ? , == , ~ , :\ or / as a character in your search word, put it in double quotes, e.g. "*"



N	Concordance	Set	Ta
1	a fortune to make, Dragon, and a beautiful young lady to make it for; and		
2	all who admire you--and all do --as a beautiful and elegant lady, know you to		
3	rude to me, and that there had been a beautiful young lady at Miss Havisham's		
4	of the fireplace, imagination seated a beautiful young lady, with a very little		
5	Polly; 'and see how quiet she is! what a beautiful little lady, ain't she?' This		
6	seen so few that I hardly know. What a beautiful, mild face that lady's is!' 'Ah!'		
7	young lady who is good, accomplished, beautiful. You are good, accomplished,		
8	of. And she says, that lady rich and beautiful that I can never come near,		
9	iron-master. "The lady was wealthy and beautiful, and had a liking for the girl, and		
10	an enclosed letter to a young and beautiful lady, then unmarried, in		
11	If the young lady had appeared beautiful by night, she was perfectly		
12	is the boofer lady?' Now, the boofer, or beautiful, lady was Bella; and whereas		
13	full in Belinda's face, as he responded--'Beautiful, indeed!' The lady cast down		
14	not, my Lady Dedlock reigns supreme. Beautiful, elegant, accomplished, and		
15	off now, however, and I answered, "The beautiful young lady at Miss Havisham's,		
16	down, and looked eagerly inside for the beautiful young lady. Alas! There was		
17	kindly on her. Fresh hope was in it. The beautiful lady who had soothed and		
18	I know. We are here, my love.' The beautiful lady released her hold of		
19	the life and adventures of the beautiful Lady Tollimglower, deceased;		
20	hope of again looking on the face of the beautiful lady. His expectations were		
21	word that had fallen from the lips of the beautiful lady, sounded to Florence like		
22	Begin by thinking well of me,' said the beautiful lady. 'Begin by believing that I		
23	bosom. There was a short silence. The beautiful lady, who at first had seemed to		
24	yet been married twelve months to the beautiful young lady I had seen in the		
25	to the great oath he had sworn to the beautiful young lady, refusing several		
26	a very uncommon sort of person this beautiful young lady must have been, to		
27	his eagerness to hear more. 'A very beautiful young lady,' said Mr Cheeryble,		
28	is living now. She is a lady and very beautiful. And I love her!" With a last faint		

In line 14, the search-word and the context-word are in separate sentences. To avoid this, specify a suitable stop as shown here:



context word(s) & context search horizons

lady

L2 L3 L4 L5 L6 to R2 R3 R4 R5 R6

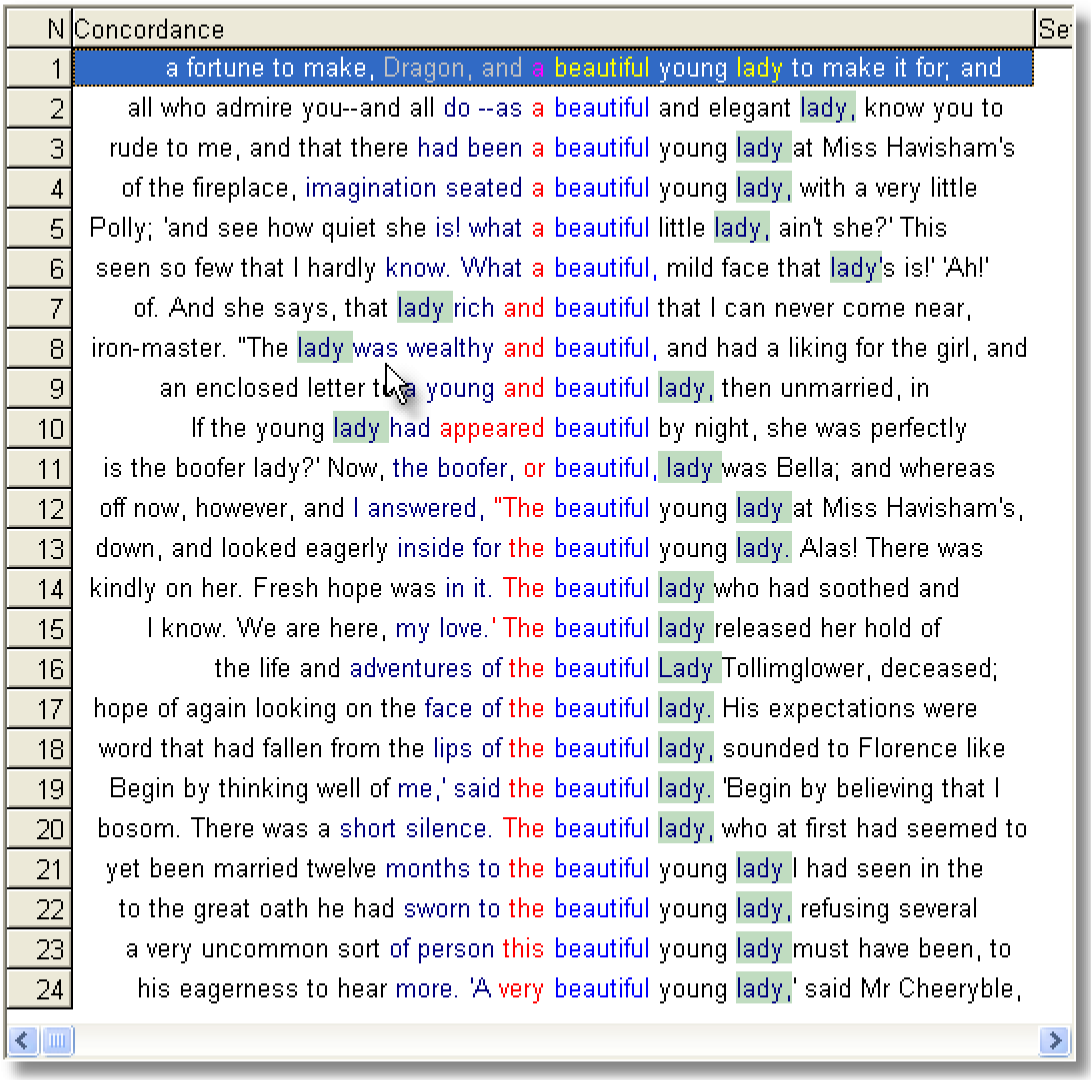
Example:

Search Word book & Context Word hotel, 5L to 5R gets book* only if hotel is found within 5 words left or right of book**

stop at sentence break

exclude if search is or context contains:

and with the same settings you will get results like these:



N	Concordance	Se
1	a fortune to make, Dragon, and a beautiful young lady to make it for; and	
2	all who admire you--and all do --as a beautiful and elegant lady, know you to	
3	rude to me, and that there had been a beautiful young lady at Miss Havisham's	
4	of the fireplace, imagination seated a beautiful young lady, with a very little	
5	Polly; 'and see how quiet she is! what a beautiful little lady, ain't she?' This	
6	seen so few that I hardly know. What a beautiful, mild face that lady's is!' 'Ah!'	
7	of. And she says, that lady rich and beautiful that I can never come near,	
8	iron-master. "The lady was wealthy and beautiful, and had a liking for the girl, and	
9	an enclosed letter to a young and beautiful lady, then unmarried, in	
10	If the young lady had appeared beautiful by night, she was perfectly	
11	is the boofer lady?' Now, the boofer, or beautiful, lady was Bella; and whereas	
12	off now, however, and I answered, "The beautiful young lady at Miss Havisham's,	
13	down, and looked eagerly inside for the beautiful young lady. Alas! There was	
14	kindly on her. Fresh hope was in it. The beautiful lady who had soothed and	
15	I know. We are here, my love.' The beautiful lady released her hold of	
16	the life and adventures of the beautiful Lady Tollimglower, deceased;	
17	hope of again looking on the face of the beautiful lady. His expectations were	
18	word that had fallen from the lips of the beautiful lady, sounded to Florence like	
19	Begin by thinking well of me,' said the beautiful lady. 'Begin by believing that I	
20	bosom. There was a short silence. The beautiful lady, who at first had seemed to	
21	yet been married twelve months to the beautiful young lady I had seen in the	
22	to the great oath he had sworn to the beautiful young lady, refusing several	
23	a very uncommon sort of person this beautiful young lady must have been, to	
24	his eagerness to hear more. 'A very beautiful young lady,' said Mr Cheeryble,	

If you have specified a context word, you can re-sort on it. Also, the context words will be in their own special colour 60.

Note: the search only takes place within the current concordance line with the number of characters defined as characters to save 211. That is, if for example you choose search horizons 25L and 25R, but only 1000 characters are saved in each line, there might not be 25 words on either side of the search-word to examine when seeking the context word or phrase if there was extensive mark-up as well.

7.13 editing concordances

The point of it...

You may well find you have got some entries which weren't what you expected. Suppose you have done a search for **SHRIMP*/PRAWN*** -- you may find a mention of *Shrimpton* in the listing. It's easy to clean up the listing by simply pressing **Del** on each unwanted line. (Do a sort on the search word first so as to get all the *Shrimptons* next to each other.) The line will turn a light grey colour.

Pressing **Ins** will restore it, if you make a mistake. To delete or restore ALL the lines from the current line to the bottom, press the grey - key or the grey + key by the numeric keypad. When you have finished marking unwanted lines, you can choose (Ctrl+Z or ⚡) to zap 129 the deleted lines.

If you're a teacher you may want to blank 168 out the search words: to do so, press the spacebar. Pressing the spacebar again will restore it, so don't worry!

7.13.1 remove duplicates

The problem

Sometimes one finds that text files contain duplicate sections, either because the corpus has become corrupted through being copied numerous times onto different file-stores or because they were not edited effectively, e.g. a newspaper has several different editions in the same file. The result can sometimes be that you get a number of repeated concordance lines.

Solution

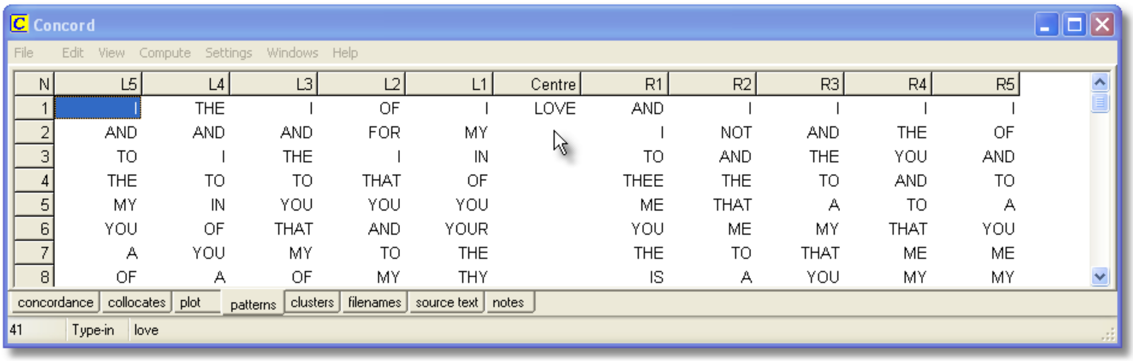
If you choose *Edit |Remove Duplicates*, **Concord** goes through your concordance lines and if it finds any two where the stored concordance lines 225 are identical, regardless of the filename, date etc. it will mark one of these for deletion. That is, it checks all the "characters to save 225" to see whether the two lines are identical. If you set this to 150 or so it is highly unlikely that false duplicates will be identified, since every single character, comma, space etc. would have to match.

Check before you zap...

At the end it will sort all the lines so you can see which ones match each other before you decide finally to zap 129 the ones you really don't want.

7.14 patterns

When you have a collocation window open, one of the tab windows shows "Patterns". This will show the collocates (words adjacent to the search word), organised in terms of frequency within each column. That is, the top word in each column is the word most frequently found in that position. The second word is the second most frequent.



Concord
File Edit View Compute Settings Windows Help
concordance collocates plot patterns clusters filenames source text notes
41 Type-in love

N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	I	THE	I	OF	I	LOVE	AND	I	I	I	I
2	AND	AND	AND	FOR	MY		I	NOT	AND	THE	OF
3	TO	I	THE	I	IN		TO	AND	THE	YOU	AND
4	THE	TO	TO	THAT	OF		THEE	THE	TO	AND	TO
5	MY	IN	YOU	YOU	YOU		ME	THAT	A	TO	A
6	YOU	OF	THAT	AND	YOUR		YOU	ME	MY	THAT	YOU
7	A	YOU	MY	TO	THE		THE	TO	THAT	ME	ME
8	OF	A	OF	MY	THY		IS	A	YOU	MY	MY

In R1 position (one word to the right of the search-word **love**) there seem to be both intimate (**thee**) and formal (**you**) pronouns associated with **love** in Shakespeare. And looking at L1 position it seems that speakers talk more of their love for another than of another's love for them.

The minimum frequency and length for one of the words to be shown at all, is the minimum frequency/length for collocates 187.

The point of it...

The effect is to make the most frequent items in the neighbourhood of the search word "float up" to the top. Like collocation, this helps you to see lexical patterns in the concordance.

You can also highlight any given pattern collocate 186 in your concordance display.

7.15 re-sorting

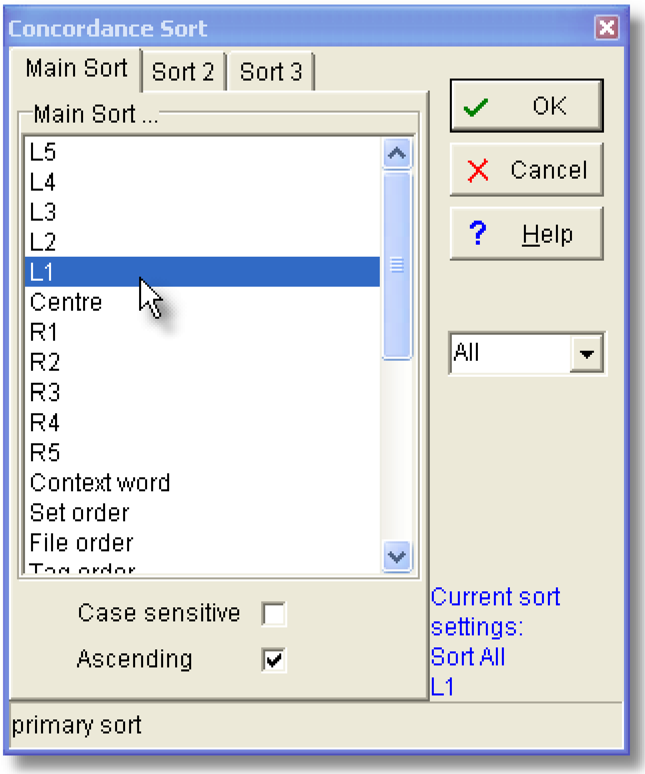
How to do it...

Sorting can be done simply by pressing the top row of any list. Or by pressing F6. Or by choosing the menu option.

The point of it...

The point of re-sorting is to find characteristic lexical patterns. It can be hard to see overall trends in your concordance lines, especially if there are lots of them. By sorting them you can separate out multiple search words and examine the immediate context to left and right. For example you may find that most of the entries have "in the" or "in a" or "in my" just before the search word -- sorting by the second word to the left of the search word will make this much clearer.
Sorting is by a given number of words to the left or right (L1 [=1 word to the left of the search word], L2, L3, L4, L5, R1 [=1 to the right], R2, R3, R4, R5), on the search word itself, the context word (if one was specified), the nearest tag 199, the distance to the nearest tag, a set category 168 of your own choice, or original file order (file).

Main Sort



Case sensitive □

Ascending ☑

Current sort settings: Sort All L1

primary sort

The listing can be sorted by three criteria at once. A Main Sort on Left 1 (L1) will sort the entries according to the alphabetical order of the word immediately to the left of the search word. A second sort (Sort 2) on R2 would re-order the listing by tie-breaking, that is: only where the L1 words (immediately to the left of the search word) matched exactly, and would place these in alphabetical order of the words 2 to the right of the search word. For very large concordances you may find the third sort (Sort 3) useful: this is an extra tie-breaker in cases where the second sort matches.

For many purposes tie-breaking is unnecessary, and will be ignored if the "activated" box is not checked.

default sort

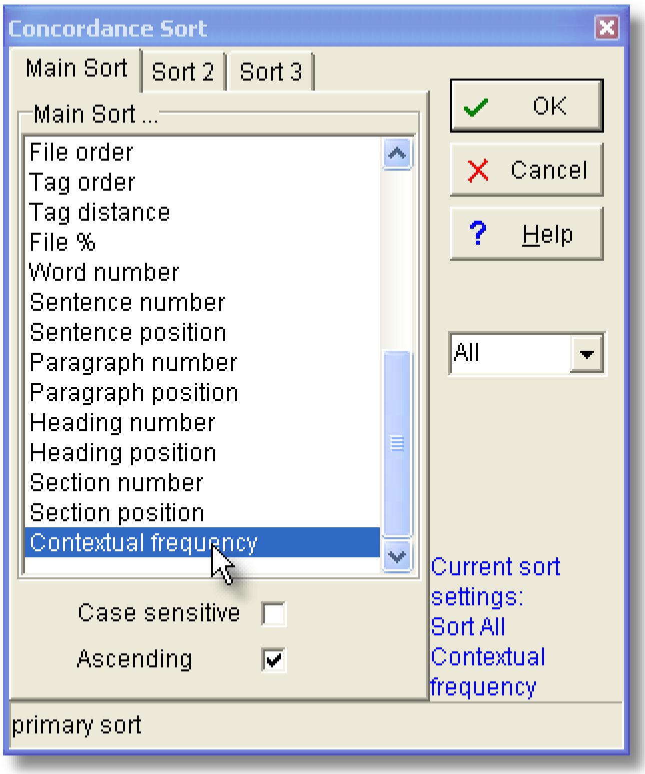
This is set in the main controller settings 220.

sorting by set (user-defined categories 168)

You can also sort by set, if you have chosen to classify the concordance lines according to your own scheme, using letters from **A** to **Z** or **a** to **z** or longer strings. The sort will put the classified lines first, in category order, followed by any unclassified lines. See Nearest Tag 199 for details of sorting by tags. See colour categories 51 for a more sophisticated way of using the Set column.

other sorts

As the screenshot below shows, you can also sort by a number of other criteria, most of these accessible simply by clicking on their column header.



Main Sort ...

File order
Tag order
Tag distance
File %
Word number
Sentence number
Sentence position
Paragraph number
Paragraph position
Heading number
Heading position
Section number
Section position
Contextual frequency

Case sensitive □

Ascending ☑

All

Current sort settings:
Sort All
Contextual frequency

primary sort

The "contextual frequency" sort means sorting on the average ranking frequency of all the words in each concordance line which don't begin with a capital letter. For this you will be asked to specify your reference corpus wordlist. The result will be to sort those lines which contain "easy" (highly frequent) words at the top of the list.

All

By default you sort all the lines; you may however type in for example 5-49 to sort those lines only.

Ascending

If this box is checked, sort order is from **A** to **Z**, otherwise it's from **Z** to **A**.

See also: WordList sort 315, KeyWords sort 252, Choosing Language 81

7.15.1 re-sorting: dispersion plot

This automatically re-sorts the dispersion plot, rotating through these options:
alphabetically (by file-name)
in *frequency* order (in terms of hits per 1,000 words of running text)
by first occurrence in the source text(s): *text order*
by *range*: the gap between first and last occurrence in the source text.

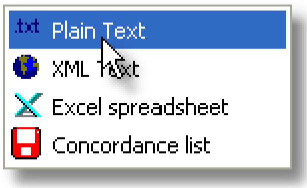
see also: Dispersion Plot 191

7.16 saving and printing

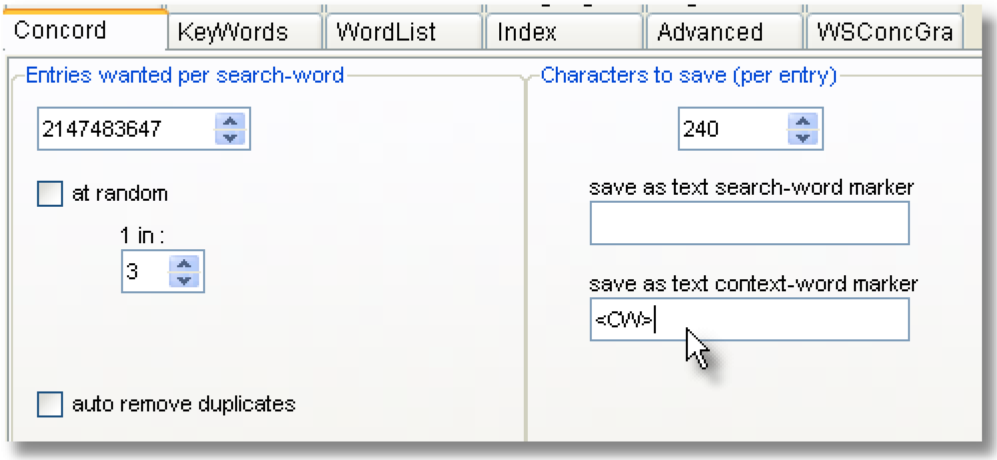
You can save the concordance (and its collocates & other dependent results if these were stored when the concordance was generated) either as a Text File (e.g. for importing into a word processor) or as a file of results which you can subsequently *Open* (in the main menu at the top) to view again at a later date. When you leave **Concord** you'll be prompted to save if you haven't already done so.

Saving a concordance allows you to return later and review collocates, dispersion plots, clusters.

You can Print 80 using the Windows printer attached to your system. You will get a chance to specify the number of pages to print. The font will approximate the one you can see on your screen. If you use a colour printer or one with various shades of grey, the screen colours will be copied to your printer. If it is a black-and-white printer, coloured items will come in *italics* if your printer can do italics.
Concord prints as much of your concordance plus associated details as your printing paper settings allow, the edges being shown in Print Preview 97.



If you choose to save as text using , and if you have (optionally) marked out the search-word and/or context word in the Controller 222 like this



Entries wanted per search-word: 2147483647

□ at random

1 in : 3

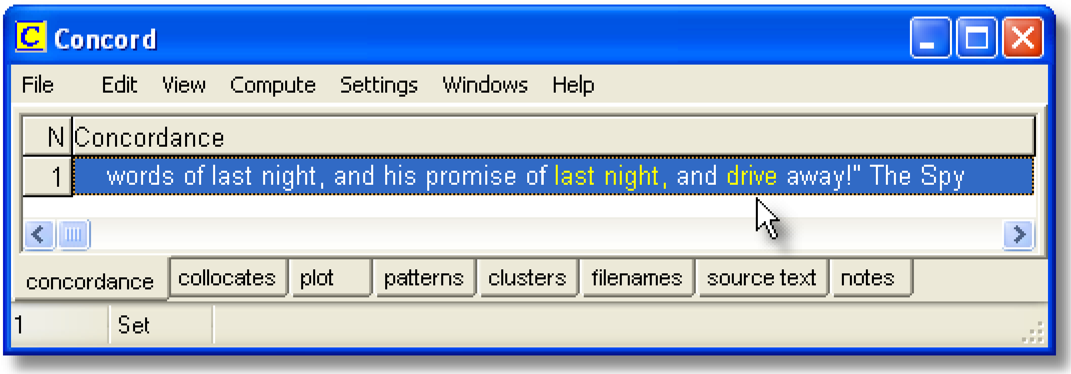
□ auto remove duplicates

Characters to save (per entry): 240

save as text search-word marker:

save as text context-word marker: <CW>

whatever you have put will get inserted in the .txt file. In the above example, doing a search through 23 Dickens texts for `last night` with `drive` as the context word, a concordance looking like this


Concord
File Edit View Compute Settings Windows Help
concordance collocates plot patterns clusters filenames source text notes
1 Set

N	Concordance
1	words of last night, and his promise of last night, and drive away!" The Spy

produced this in the txt file:

```
rry, tell him yourself to give him no restorative but air, and to
remember my words of last night, and his promise of last night, and
<CW>drive away!" The Spy withdrew, and Carton seated himself at the
table, resting his forehead on his h
```

See also : using the clipboard 422 to get a concordance into Word or another application.

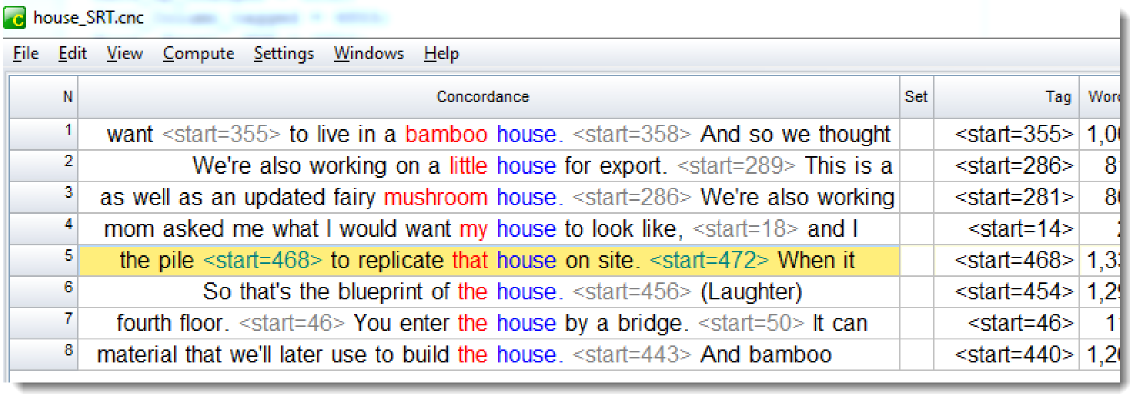
7.17 sounds & video

The point of it

Suppose you do a concordance of "elephant" and want to hear how the word is actually spoken in context. Is the last vowel a schwa? Does the second vowel sound like "i" or "e" or "u" or a schwa?

How to do it...

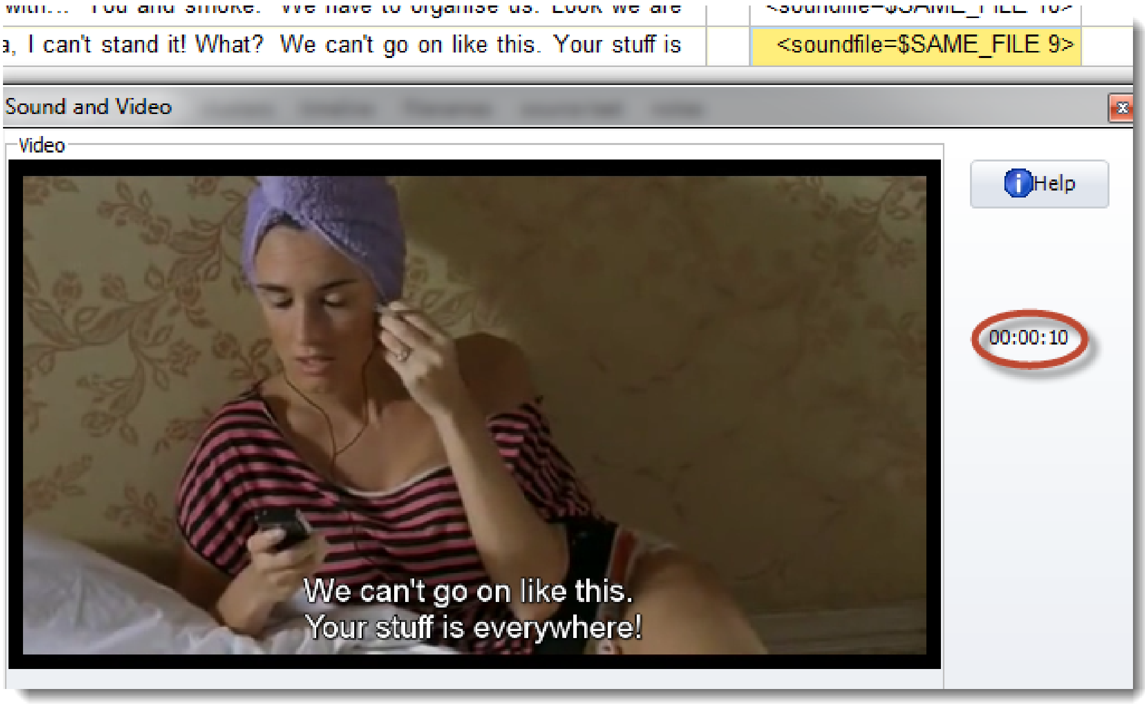
If you have defined tags which refer to multimedia files, and if there are any such tags in the "tag-context" of a given concordance line, you can hear or see the source multimedia. The tag will be parsed 147 to identify the file needed, if necessary downloading it from a web address, and then played.



N	Concordance	Set	Tag	Wor
1	want <start=355> to live in a bamboo house. <start=358> And so we thought		<start=355>	1,0
2	We're also working on a little house for export. <start=289> This is a		<start=286>	8
3	as well as an updated fairy mushroom house. <start=286> We're also working		<start=281>	8
4	mom asked me what I would want my house to look like, <start=18> and I		<start=14>	
5	the pile <start=468> to replicate that house on site. <start=472> When it		<start=468>	1,3
6	So that's the blueprint of the house. <start=456> (Laughter)		<start=454>	1,2
7	fourth floor. <start=46> You enter the house by a bridge. <start=50> It can		<start=46>	1
8	material that we'll later use to build the house. <start=443> And bamboo		<start=440>	1,2

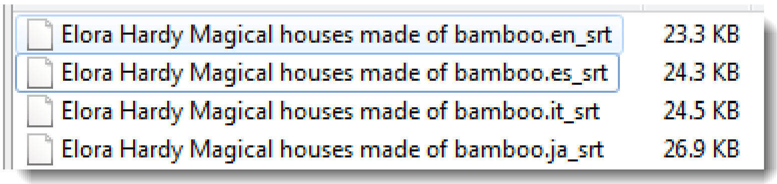
In this screenshot we see a concordance where there is a tag inserted periodically in the text file. To play the media file,choose *File* | *Play media file*, or double-click the *Tag* column.

Video files can be played if the free VLC Media Player is installed (see http://www.vlcapp.com/). The next screenshot below shows a concordance line with, in the *Nearest Tag* column, the mark-up saying that the source text and the video file have the same file-name (except that the latter ends .AVI and the former .TXT). A double-click on the Tag (yellow highlighted cell) brought up the video screen you can see below,


a, I can't stand it! What? We can't go on like this. Your stuff is <soundfile=$SAME_FILE 9>

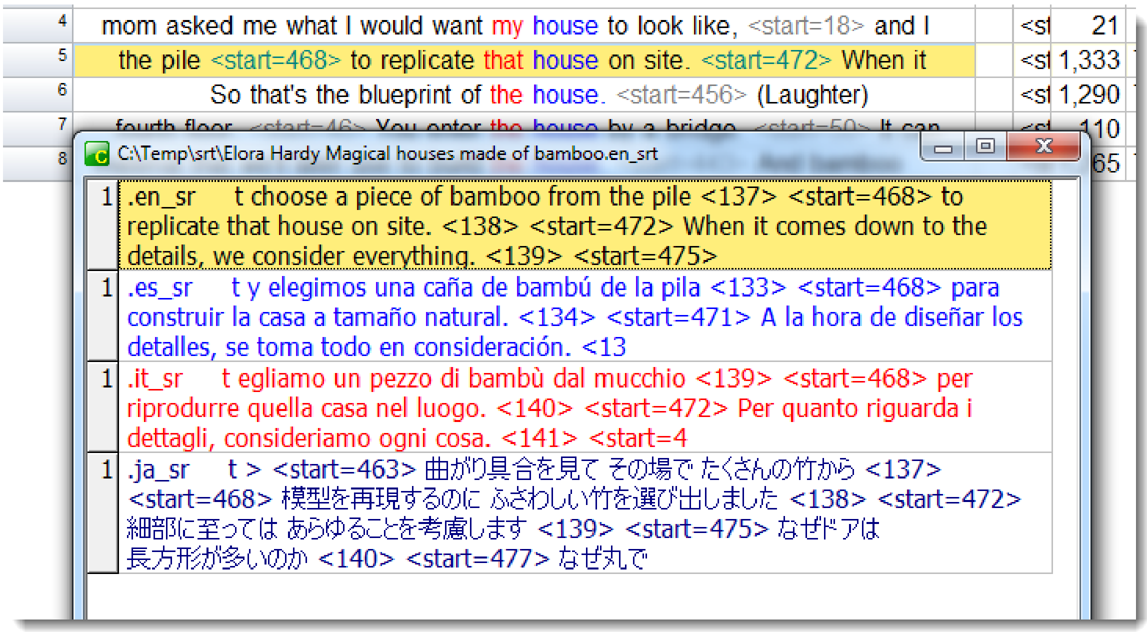
and that has now played to the tenth second, then paused. You can see in the case of this particular video that there is a sub-title with the same words that are in the concordance above (though there is no guarantee you will see sub-titles for all videos).

If you build up a collection of TED talks like these where the same video in English has transcripts in several languages,



Elora Hardy Magical houses made of bamboo.en_srt	23.3 KB
Elora Hardy Magical houses made of bamboo.es_srt	24.3 KB
Elora Hardy Magical houses made of bamboo.it_srt	24.5 KB
Elora Hardy Magical houses made of bamboo.ja_srt	26.9 KB

you can get to see the different translations:



4 mom asked me what I would want my house to look like, <start=18> and I <st 21

5 the pile <start=468> to replicate that house on site. <start=472> When it <st 1,333

6 So that's the blueprint of the house. <start=456> (Laughter) <st 1,290

7 fourth floor. <start=46> You enter the house by a bridge. <start=50> It can <st 110

8 65

C:\Temp\srt\Elora Hardy Magical houses made of bamboo.en_srt

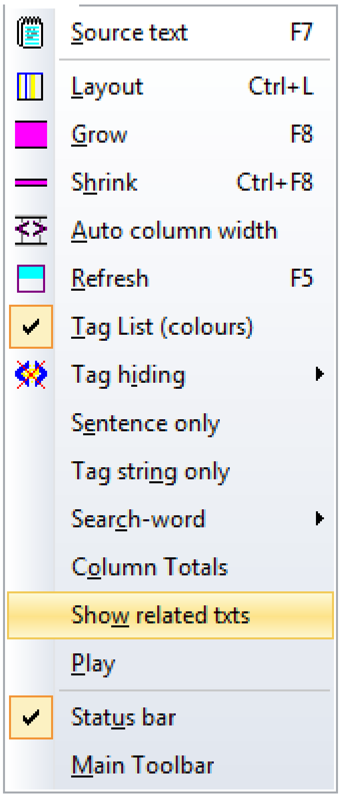
1 .en_sr t choose a piece of bamboo from the pile <137> <start=468> to replicate that house on site. <138> <start=472> When it comes down to the details, we consider everything. <139> <start=475>

1 .es_sr t y elegimos una caña de bambú de la pila <133> <start=468> para construir la casa a tamaño natural. <134> <start=471> A la hora de diseñar los detalles, se toma todo en consideración. <13

1 .it_sr t egliamo un pezzo di bambù dal mucchio <139> <start=468> per riprodurre quella casa nel luogo. <140> <start=472> Per quanto riguarda i dettagli, consideriamo ogni cosa. <141> <start=4

1 .ja_sr t > <start=463> 曲がり具合を見て その場で たくさんの竹から <137> <start=468> 模型を再現するのに ふさわしい竹を選び出しました <138> <start=472> 細部に至っては あらゆることを考慮します <139> <start=475> なぜドアは 長方形が多いのか <140> <start=477> なぜ丸で

by choosing *View* | *Show related txts* in the menu.



See also: Multi-media Tag syntax 147, Obtaining Sound and Video files 215, Handling Tags 132, Making a Tag File 141, Showing Nearest Tags in Concord 199, Tag Concordancing 198, Types of Tag 145, Viewing the Tags 390, Using Tags as Text Selectors 134, Tags in WordList 311

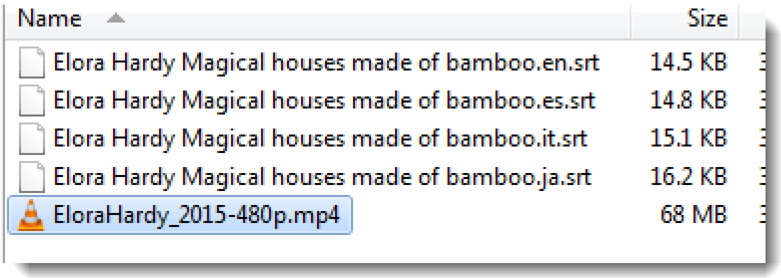
7.17.1 obtaining sound and video files

Sources of sound and video files

WordSmith does not provide or include corpora. However, there are specialised corpora such as NECTE, MICASE, ICE and then there are publicly available sources such as the TED Talks. You are expected to respect copyright provisions in all cases.

There is a lot of useful advice at
TED Open Translation Project where you will find transcripts.

These text files in English (.en), Spanish (.es), Italian (.it) and Japanese (.ja) were downloaded from there and later converted using the Text Converter 374



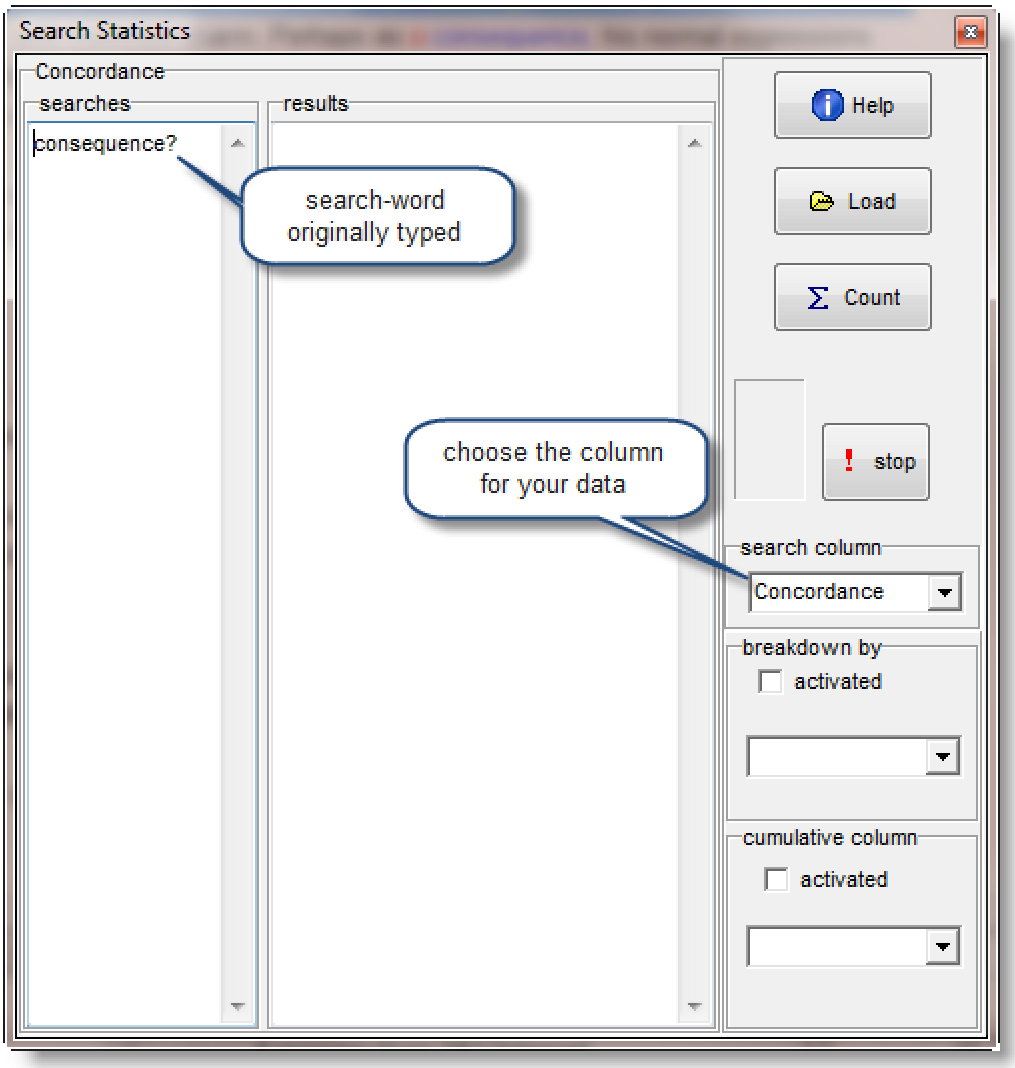
Name	Size
Elora Hardy Magical houses made of bamboo.en.srt	14.5 KB
Elora Hardy Magical houses made of bamboo.es.srt	14.8 KB
Elora Hardy Magical houses made of bamboo.it.srt	15.1 KB
Elora Hardy Magical houses made of bamboo.ja.srt	16.2 KB
EloraHardy_2015-480p.mp4	68 MB

If you wish to use a transcript and sound file format which is incompatible with the syntax 147 described here, please contact 425 us.

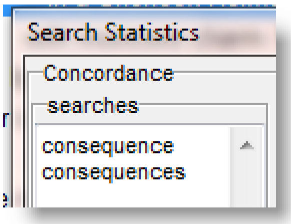
7.18 summary statistics

The idea is to be able to break down your concordance data. For example, you've just done a concordance of **consequence?** which has given you lots of singulars and lots of plurals and you want to know how many there are of each.

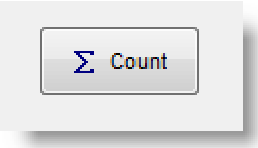
Choose *Summary Statistics* in the *Compute* menu.



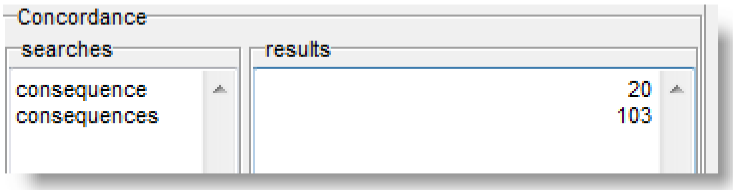
The *searches* window will at first contain a copy of what you typed in when you created the concordance. To distinguish between singular and plural, change that to



and press Count;



assuming that *search column* has *Concordance* selected, you will get something like this:



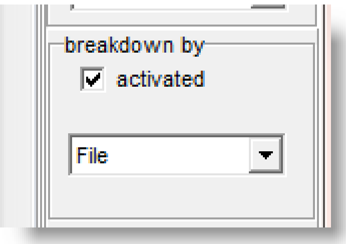
consequence 20
consequences 103

Advanced Summary Statistics features

Breakdown

The idea here is to be able to break down your results further, using another category in your existing concordance data, such as the files the data came from. In our example, we might want to know for **consequence** and **consequences**, how many of the text files contained each of the two forms.

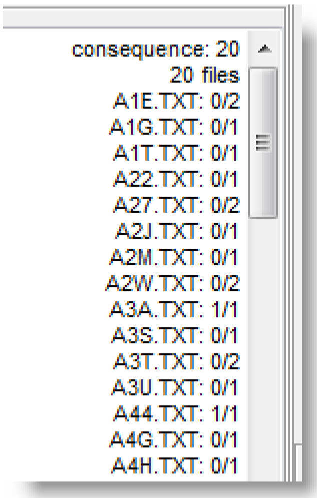
To generate the breakdown, activate it and choose the category you need.



☑ activated

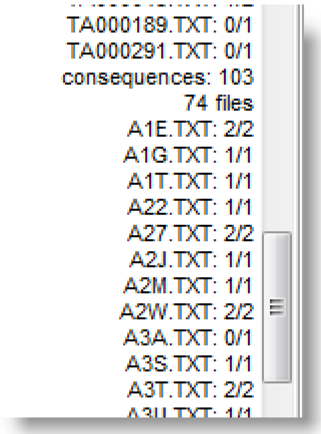
File

The results window will now show something like this



```
consequence: 20
       20 files
 A1E.TXT: 0/2
 A1G.TXT: 0/1
 A1T.TXT: 0/1
 A22.TXT: 0/1
 A27.TXT: 0/2
 A2J.TXT: 0/1
 A2M.TXT: 0/1
 A2W.TXT: 0/2
 A3A.TXT: 1/1
 A3S.TXT: 0/1
 A3T.TXT: 0/2
 A3U.TXT: 0/1
 A44.TXT: 1/1
 A4G.TXT: 0/1
 A4H.TXT: 0/1
```

where it is clear that the singular **consequence** came 20 times in 20 different files, the first being file **A3A.TXT**. Further down you will find the results for **consequences**:



TA000189.TXT: 0/1
TA000291.TXT: 0/1
consequences: 103
74 files
A1E.TXT: 2/2
A1G.TXT: 1/1
A1T.TXT: 1/1
A22.TXT: 1/1
A27.TXT: 2/2
A2J.TXT: 1/1
A2M.TXT: 1/1
A2W.TXT: 2/2
A3A.TXT: 0/1
A3S.TXT: 1/1
A3T.TXT: 2/2

which appeared 103 times in 74 files, and that in the first of these, **A1E.TXT**, it came twice.

Cumulative column

see the explanation for WordList 304

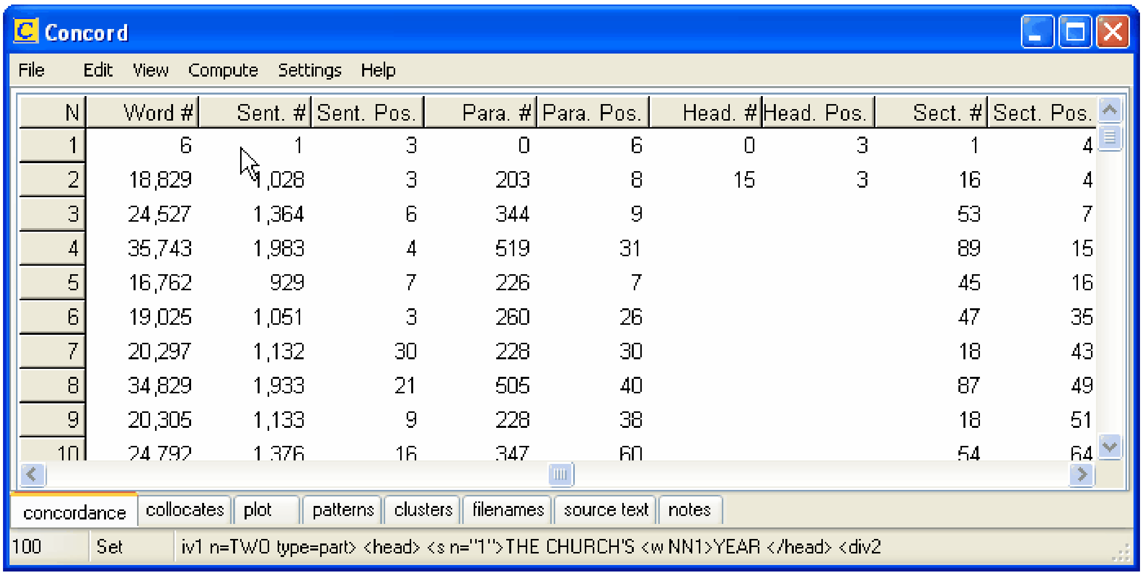
Load button

see the explanation for count data frequencies 66.

7.19 text segments in Concord

A concordance line brings with it information about which segment of the text it was found in.

In the screenshot below, a concordance on **year** was carried out; the listing has been sorted by Heading Position -- in the top 2 lines, **year** is found as the 3rd word of a heading. The advantage of this is that it is possible to identify search-words occurring near sentence starts, near the beginning of sections, of headings, of paragraphs.



N	Word #	Sent. #	Sent. Pos.	Para. #	Para. Pos.	Head. #	Head. Pos.	Sect. #	Sect. Pos.
1	6	1	3	0	6	0	3	1	4
2	18,829	1,028	3	203	8	15	3	16	4
3	24,527	1,364	6	344	9			53	7
4	35,743	1,983	4	519	31			89	15
5	16,762	929	7	226	7			45	16
6	19,025	1,051	3	260	26			47	35
7	20,297	1,132	30	228	30			18	43
8	34,829	1,933	21	505	40			87	49
9	20,305	1,133	9	228	38			18	51
10	24,792	1,376	16	347	60			54	64

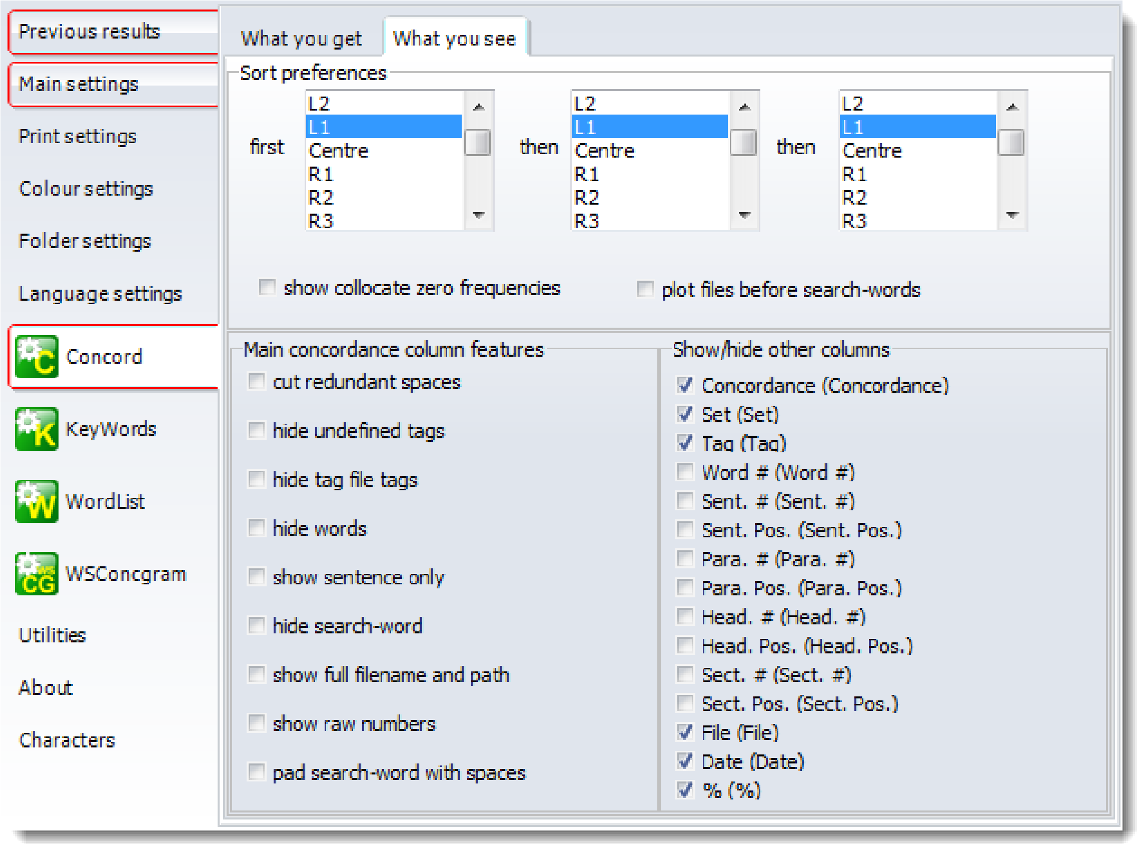
100 Set iv1 n=TWO type=part> <head> <s n="1">THE CHURCH'S <w NN1>YEAR </head> <div2

You can toggle the numbers between raw numbers and percentages 220.

See also: Start and end of text segments 146.

7.20 viewing options

Access these options in the main Controller, via *Concord* | *What you see*.



Sort preferences

first: L2, **L1**, Centre, R1, R2, R3 — then: L2, **L1**, Centre, R1, R2, R3 — then: L2, **L1**, Centre, R1, R2, R3

□ show collocate zero frequencies □ plot files before search-words

Main concordance column features

- □ cut redundant spaces
- □ hide undefined tags
- □ hide tag file tags
- □ hide words
- □ show sentence only
- □ hide search-word
- □ show full filename and path
- □ show raw numbers
- □ pad search-word with spaces

Show/hide other columns

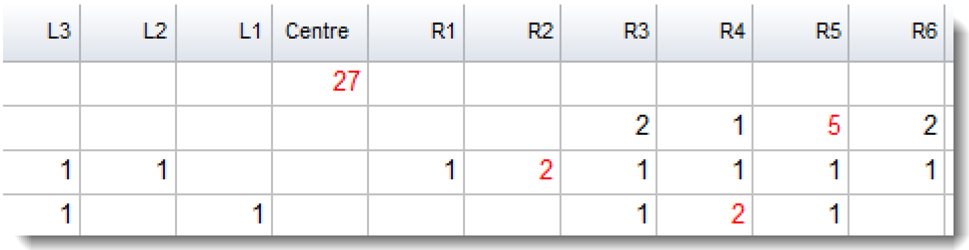
- ☑ Concordance (Concordance)
- ☑ Set (Set)
- ☑ Tag (Tag)
- □ Word # (Word #)
- □ Sent. # (Sent. #)
- □ Sent. Pos. (Sent. Pos.)
- □ Para. # (Para. #)
- □ Para. Pos. (Para. Pos.)
- □ Head. # (Head. #)
- □ Head. Pos. (Head. Pos.)
- □ Sect. # (Sect. #)
- □ Sect. Pos. (Sect. Pos.)
- ☑ File (File)
- ☑ Date (Date)
- ☑ % (%)

Sort preferences

By default, **Concord** will sort a new concordance by the word to the left (L1), but you can set this to different values if you like. For further details, see Sorting a Concordance 208.

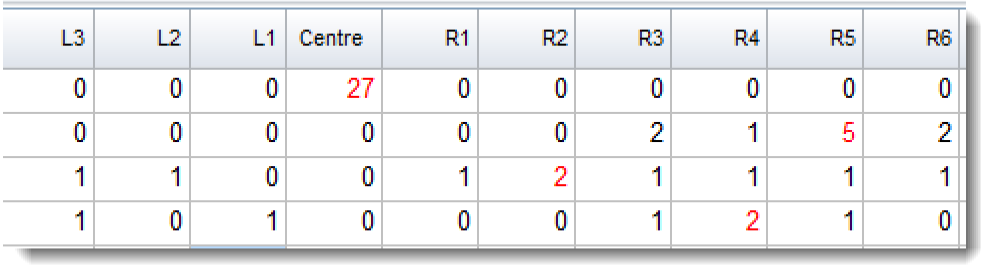
Show collocate zero frequencies

This toggles whether 0 or a blank (the default) is shown if a collocate frequency is zero.



L3	L2	L1	Centre	R1	R2	R3	R4	R5	R6
			27						
						2	1	5	2
1	1			1	2	1	1	1	1
1		1				1	2	1	

or



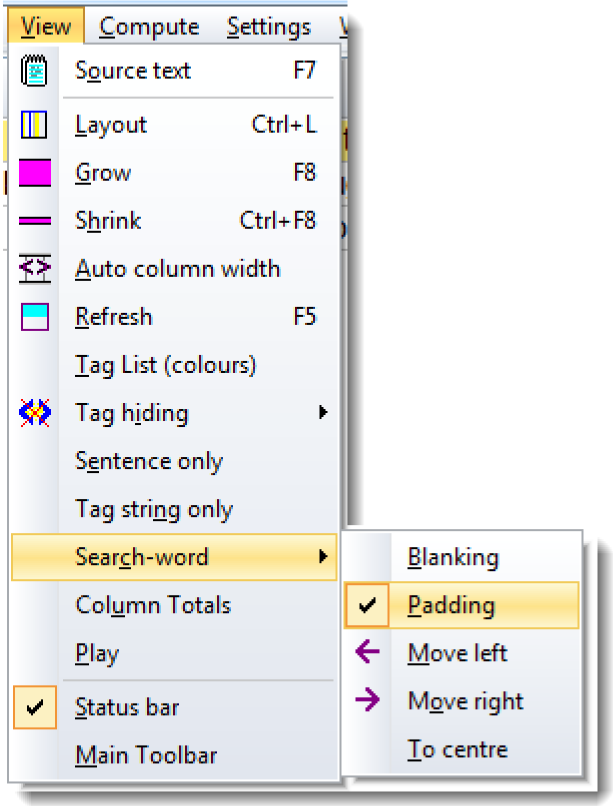
L3	L2	L1	Centre	R1	R2	R3	R4	R5	R6
0	0	0	27	0	0	0	0	0	0
0	0	0	0	0	0	2	1	5	2
1	1	0	0	1	2	1	1	1	1
1	0	1	0	0	0	1	2	1	0

Concordance View

You can choose different ways of seeing the data, and a whole set of choices as to what columns you want to display for each new concordance. You can re-instate any later if you wish by changing the Layout 87.

show full filename and path = sometimes you need to see the whole path but usually the filename alone will suffice.
cut redundant spaces = remove any double spaces
show sentence only = show the context only up to its left and right sentence boundaries 425
tag string only = show only context within two tag_string_only tags
show raw numbers = show the raw data instead of percentages e.g. for sentence position 218
hide search-word = blank it out eg. to make a guess-the-word exercise 168
pad search-word with spaces = insert a space to left and right of the search-word so it stands out better
hide undefined tags = hide those not defined in your tag file 141
hide tag file tags = hide all tags including undefined ones
hide words = show only the tags

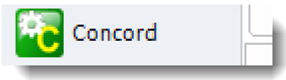
Some of the options are visible here:



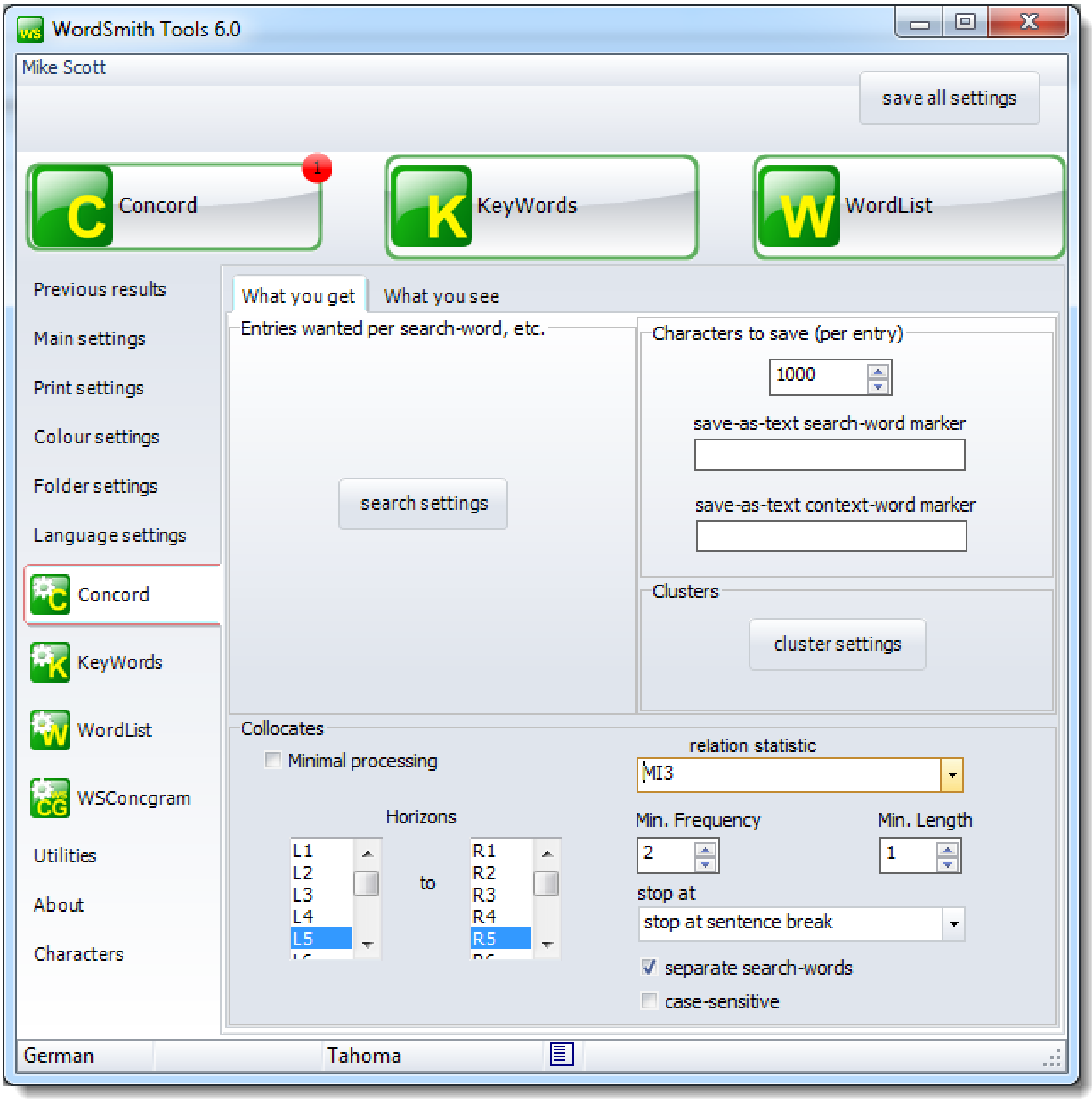
for example the sub-set visible shows an opportunity to blank out the search-word, to pad it with a space left & right, to shift the search-word left or right.

See also: Controller 222 *What you get* 222 choices 222, showing nearest tags 199, blanking out 168 the search-word, viewing more context, growing/shrinking concordance lines 166.

7.21 WordSmith controller: Concord: settings



These are found in the main Controller 4 marked *Concord.*
This is because some of the choices -- e.g. collocation horizons 180 -- may affect other Tools.



Characters to save (per entry): 1000

save-as-text search-word marker:

save-as-text context-word marker:

□ Minimal processing

relation statistic: MI3

Horizons: L5 to R5

Min. Frequency: 2

Min. Length: 1

stop at: stop at sentence break

☑ separate search-words

□ case-sensitive

When you have computed a concordance, the Concord button will have a red number (showing how many Concord windows are in use) and at the bottom of the screen you will see an icon (☰). Click that to see the list of files and their features.

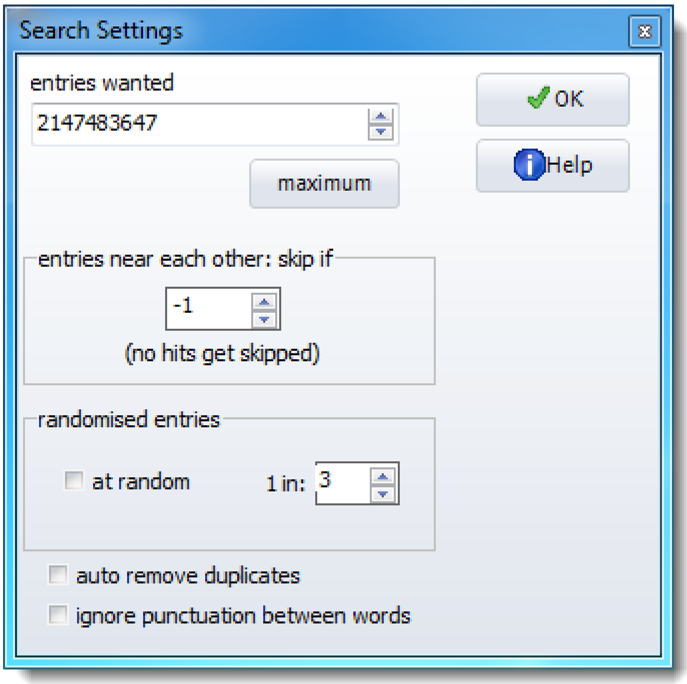
WHAT YOU GET and WHAT YOU SEE

There are 2 tabs for settings affecting *What you get* in the concordance and *What you see* in the display. There is a screenshot at Concord: viewing options 220 showing the options under *What you see*.

WHAT YOU GET

Search Settings

The search settings button lets you choose these settings:



Search Settings

entries wanted
2147483647

entries near each other: skip if
-1
(no hits get skipped)

randomised entries
□ at random 1 in: 3

□ auto remove duplicates
□ ignore punctuation between words

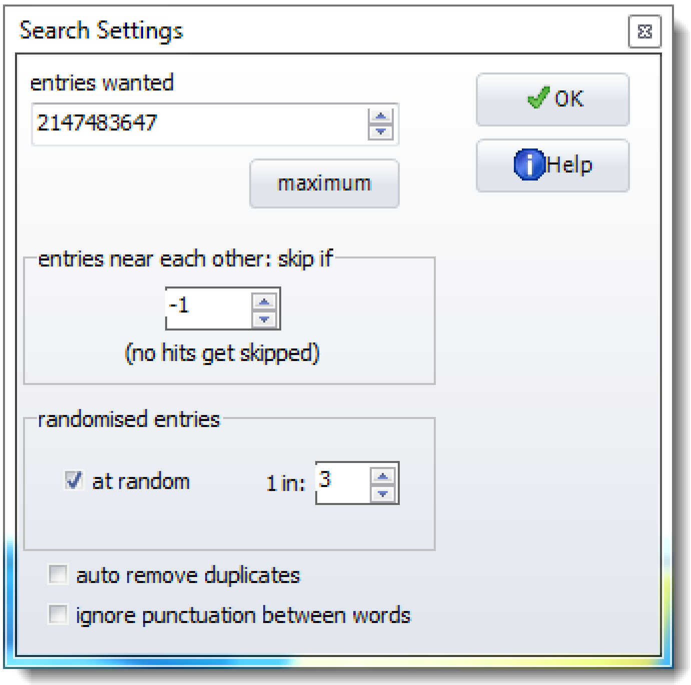
Entries Wanted

The maximum is more than 2 billion lines. This feature is useful if you're doing a number of searches and want, say, 100 examples of each. The 100 entries will be the *first 100* found in the texts you have selected. If you search for more than 1 search-word (eg. **`book/ paperback`**), you will get 100 of **`book`** and 100 of **`paperback`**.

entries near each other

allows you to force Concord to skip hits which are too close to each other. If for example you set this to 0 or 1 and your text contains
`... a lovely lovely day`
then you will only get the first of these cases if searching for **`lovely`**. The default here is -1. (If you set it to 0 then you are only allowing one hit within any given word.)

randomised entries: this feature allows you to randomise the search. Here **Concord** goes through the text files and gets the 100 entries by giving each hit a random chance of being selected. To get 100 entries **Concord** will have to have found around 450-550 hits with the settings shown below. You can set the randomiser anywhere from 1 in 2 to 1 in 1,000. See also: reduce to N 70.



Search Settings

entries wanted: 2147483647

entries near each other: skip if: -1 (no hits get skipped)

randomised entries: ☑ at random 1 in: 3

□ auto remove duplicates

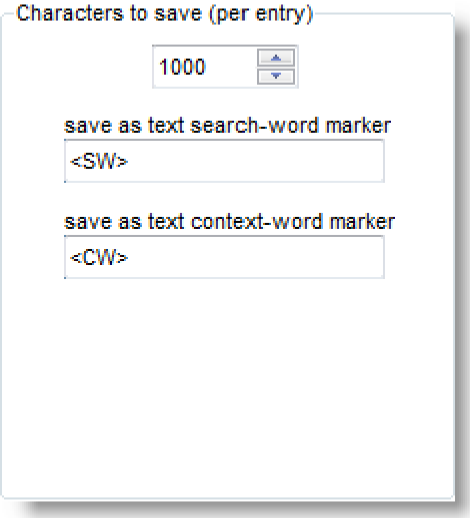
□ ignore punctuation between words

auto remove duplicates: removes any lines where the whole concordance entry matches another. (This can happen if you have a corpus where news stories get re-published in different editions by different newspapers.)

Ignore punctuation between words: this allows a search for **`BY ITSELF`** to succeed where the text contains ...**`went by, itself`**

Characters to save

Here is where you set how many characters in a concordance line will be stored as text as the concordance is generated. The default and minimum is 1000. This number of characters will be saved when you save your results, so even if you subsequently delete the source text file you can still see some context. If you grow the lines more text will be read in (and stored) as needed. There are examples here 422.



Characters to save (per entry)

1000

save as text search-word marker

<SW>

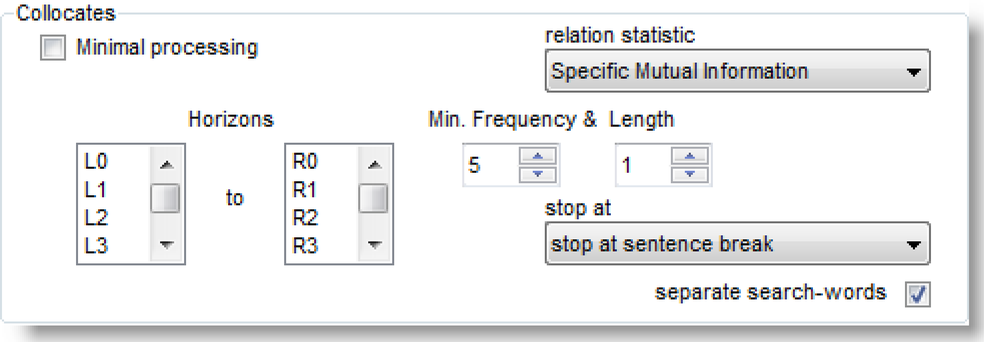
save as text context-word marker

<CW>

Save as text search-word or context-word marker: here you can also specify markers for your search-word and context-word 211.

Collocates

By default, **Concord** will compute collocates as well as the concordance, but you can set it not to if you like (*Minimal processing*). For further details, see Collocate Horizons 180 or Collocation 179



Collocates

□ Minimal processing

relation statistic: Specific Mutual Information

Horizons: L0 L1 L2 L3 to R0 R1 R2 R3

Min. Frequency & Length: 5 1

stop at: stop at sentence break

separate search-words ☑

The minimum frequency and length refer to the collocates to be shown in your listing. With the settings above, only collocates which occur at least 5 times and contain at least 1 character will be shown as long as they don't cross sentence boundaries 188.

If *separate search words* is checked and you have multiple search-terms, then you get collocates distinguishing between the different search-terms. If you want them amalgamated, clear this check-box.

Collocates relation statistic

Choose between Specific Mutual Information, MI3, Z Score, Log Likelihood. See Mutual Information Display 289 for examples of how these can differ.

WHAT YOU SEE

The options are explained at Concord: viewing options 220.

Columns

The *Columns to show/hide* list offers all the standard columns: you may uncheck ones you normally do not wish to see. This will only affect newly computed KeyWords data: earlier data uses the column visibility, size, colours etc already saved. They can be altered using the Layout 87 menu option at any time.

See also: Concord Saving and Printing 211, Concord Help Contents 158, Collocation Settings 187.

KeyWords

Section VIII

8 KeyWords

8.1 purpose

This is a program for identifying the "key" words in one or more texts. Key words are those whose frequency is unusually high in comparison with some norm. Click here for an example 235.

The point of it…

Key-words provide a useful way to characterise a text or a genre. Potential applications include: language teaching, forensic linguistics, stylistics, content analysis, text retrieval.

The program compares two pre-existing word-lists, which must have been created using the WordList tool. One of these is assumed to be a large word-list which will act as a reference file. The other is the word-list based on one text which you want to study.

The aim is to find out which words characterise the text you're most interested in, which is automatically assumed to be the smaller of the two texts chosen. The larger will provide background data for reference comparison.

Key-words and links 247 between them can be plotted 251, made into a database 237, and grouped according to their associates 241.

Online step-by-step guide showing how

8.2 index

Explanations

Settings and Procedures

Tips

Definitions

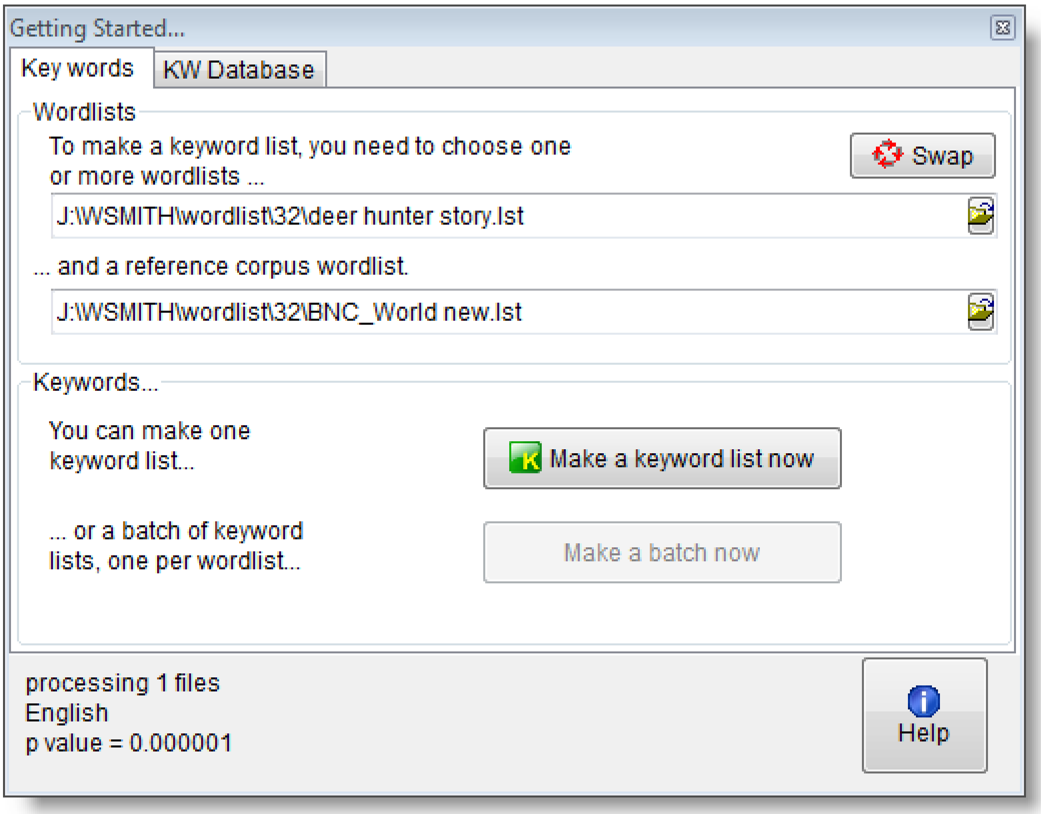
See also : WordSmith Main Index 2

8.3 ordinary two word-list analysis

The usual kind of **KeyWords** analysis. It compares the one text file (or corpus) you're chiefly interested in, with a reference corpus based on a lot of text. In the screenshot below we are interested in the key words of **`deer hunter story`** and we're using **`BNC`** as the reference corpus to compare with.

Choose Word Lists

In the dialogue box you will choose 2 files. The text file in the box above and the reference corpus file in the box below.



Getting Started...

Key words | KW Database

Wordlists

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

J:\WSMITH\wordlist\32\deer hunter story.lst

... and a reference corpus wordlist.

J:\WSMITH\wordlist\32\BNC_World new.lst

Keywords...

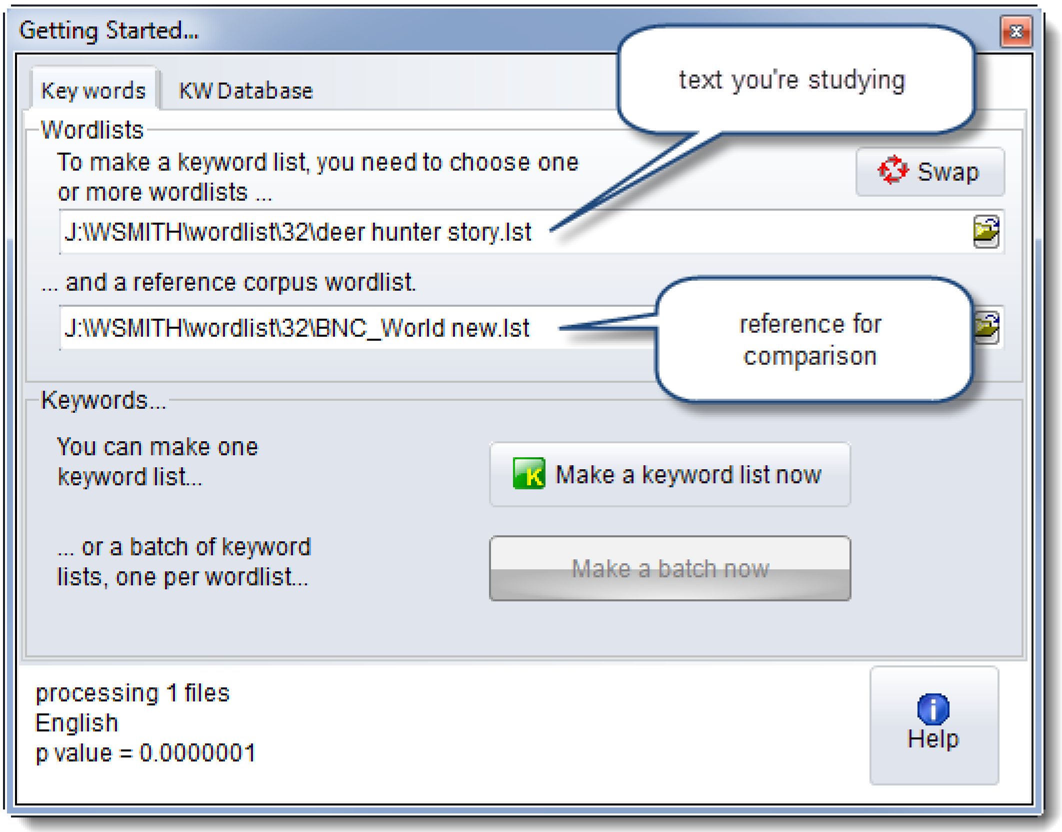
You can make one keyword list...

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

processing 1 files
English
p value = 0.000001

See also How Key Words are Calculated 245, KeyWords Settings 254

8.4 choosing files



Wordlists

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

J:\WSMITH\wordlist\32\deer hunter story.lst

... and a reference corpus wordlist.

J:\WSMITH\wordlist\32\BNC_World new.lst

Keywords...

You can make one keyword list...

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

processing 1 files
English
p value = 0.0000001

Current Text word list

In the upper box, choose a word list file.

To choose more than 1 word list file, press Control as you click to select non-adjacent lists, or Shift to select a range.

This box determines which word-list(s) you're going to find the key words of.

Reference Corpus word list

The the box below, you choose your Reference Corpus 447 List. (This can be set permanently in the main Controller Settings).

No word lists visible

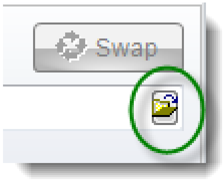
If you can't see any word lists in the displays, either change folders until you can, or go back to the WordList tool and make up at least 2 word lists: this procedure requires at least two before it can make a comparison.

Swap

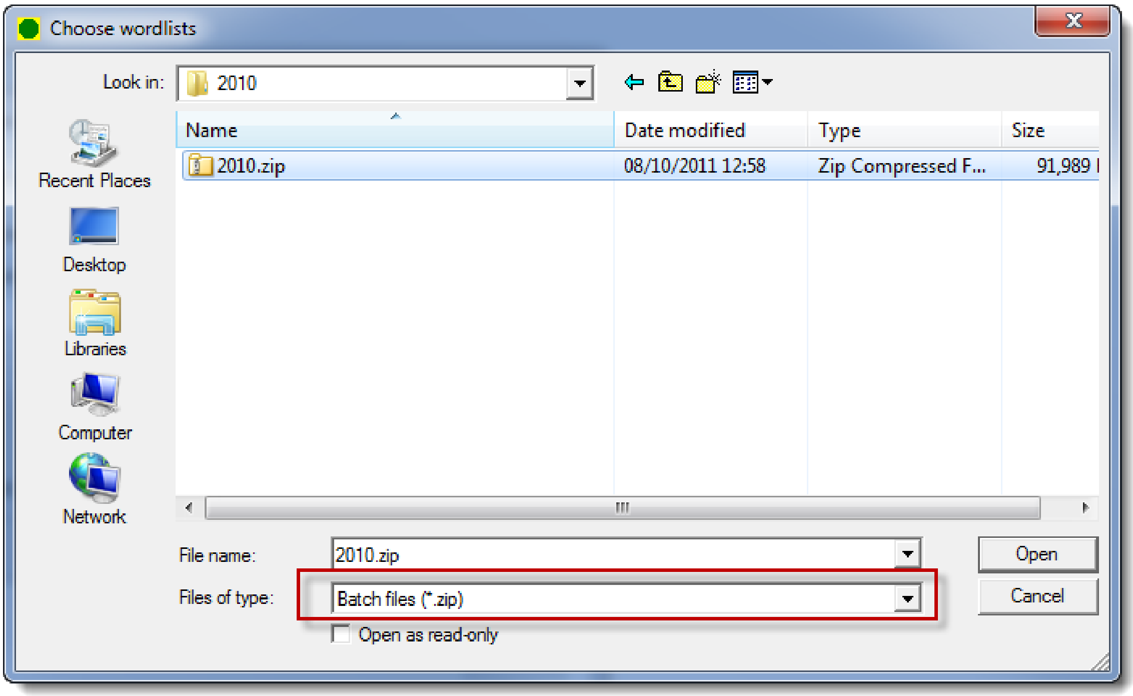
The text you're studying must be at the top. If you get them wrong, exchange them.

Advanced: working with a batch file

Click the browse button:



and choose the batch **`.zip`** file



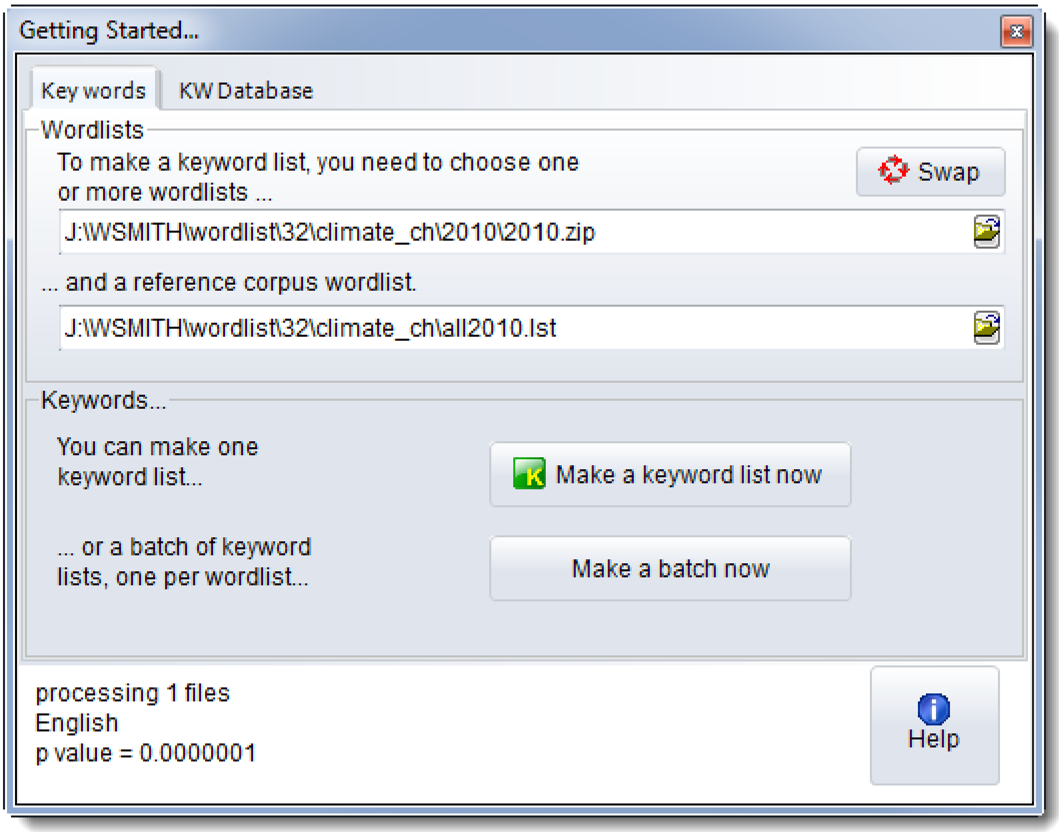
Name	Date modified	Type	Size
2010.zip	08/10/2011 12:58	Zip Compressed F...	91,989 I

File name: 2010.zip

Files of type: Batch files (*.zip)

□ Open as read-only

and we are ready to make a batch: that one 2010.zip contains many thousands of word lists.



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

J:\WSMITH\wordlist\32\climate_ch\2010\2010.zip

... and a reference corpus wordlist.

J:\WSMITH\wordlist\32\climate_ch\all2010.lst

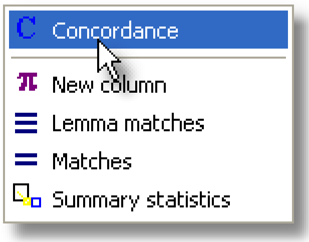
You can make one keyword list...

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

processing 1 files
English
p value = 0.0000001

8.5 concordance

With a key word or a word list list on your screen, you can choose *Compute* and



to call up a concordance of the currently selected word(s). The concordance will search for the same word in the original text file that your key word list came from.

The point of it...

is to see these same words in their original contexts.

8.6 example of key words

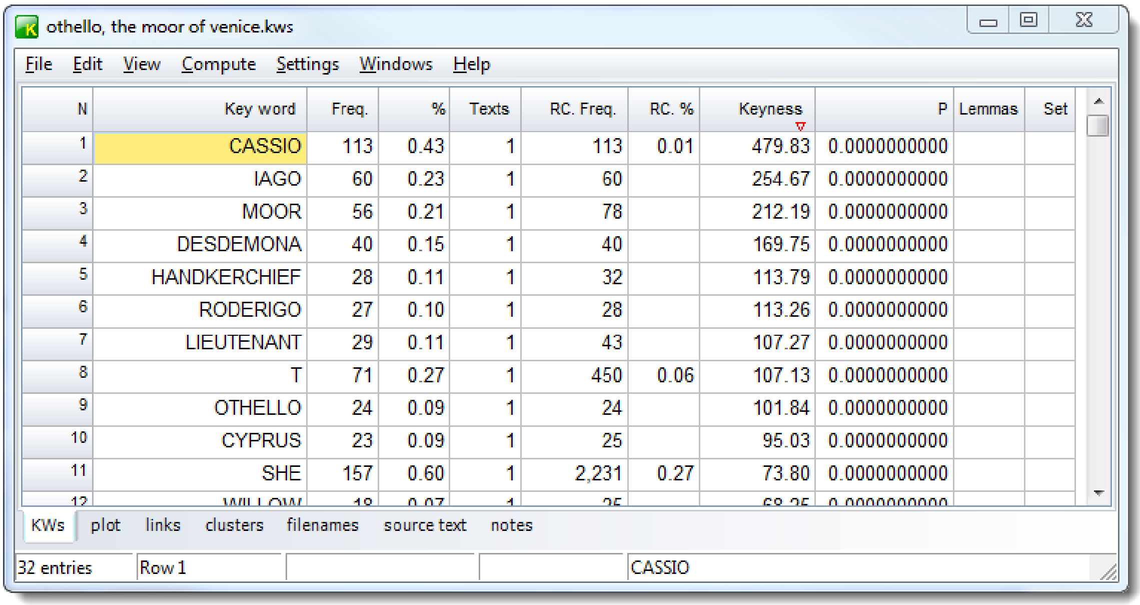
You have a collection of assorted newspaper articles. You make a word list based on these articles, and see that the most frequent word is *the.* Among the rather infrequent words in the list come examples like *hopping*, *modem, squatter, grateful*, etc.

You then take from it a 1,000 word article and make a word list of that. Again, you notice that the most frequent word is *the*. So far, not much difference.

You then get **KeyWords** to analyse the two word lists. **KeyWords** reports that the most "key" words are: *squatter, police, breakage, council, sued, Timson, resisted, community*.

These "key" words are not the most frequent words (which are those like *the*) but the words which are most unusually frequent in the 1,000 word article. Key words usually give a reasonably good clue to what the text is about.

Here is an example from the play Othello.



N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness	P	Lemmas	Set
1	CASSIO	113	0.43	1	113	0.01	479.83	0.0000000000		
2	IAGO	60	0.23	1	60		254.67	0.0000000000		
3	MOOR	56	0.21	1	78		212.19	0.0000000000		
4	DESDEMONA	40	0.15	1	40		169.75	0.0000000000		
5	HANDKERCHIEF	28	0.11	1	32		113.79	0.0000000000		
6	RODERIGO	27	0.10	1	28		113.26	0.0000000000		
7	LIEUTENANT	29	0.11	1	43		107.27	0.0000000000		
8	T	71	0.27	1	450	0.06	107.13	0.0000000000		
9	OTHELLO	24	0.09	1	24		101.84	0.0000000000		
10	CYPRUS	23	0.09	1	25		95.03	0.0000000000		
11	SHE	157	0.60	1	2,231	0.27	73.80	0.0000000000		
12	WILLOW	18	0.07	1	25		68.25	0.0000000000		

32 entries | Row 1 | CASSIO

See also: word-lists with tags as prefix 315.

8.7 keyness

8.7.1 p value

(Default=0.000001)

The ***p*** value is that used in standard chi-square and other statistical tests. This value ranges from 0 to 1. A value of .01 suggests a 1% danger of being wrong in claiming a relationship, .05 would give a 5% danger of error. In the social sciences a 5% risk is usually considered acceptable.

In the case of key word analyses, where the notion of risk is less important than that of selectivity, you may often wish to set a comparatively low ***p*** value threshold such as 0.000001 (one in 1 million) (1E-6 in scientific notation) so as to obtain fewer key words. Or you can set a low "maximum wanted" number in the main Controller 4, under *KeyWords Settings*.

If the chi-square procedure 245 is used, the computed p value will only be shown if all appropriate statistical requirements are met (all expected values >= 5).

See also: Definitions 425, choosing a reference corpus 237

8.7.2 key-ness definition

The term "key word", though it is in common use, is not defined in Linguistics. This program identifies key words on a mechanical basis by comparing patterns of frequency. (A human being, on the other hand, may choose a phrase or a superordinate as a key word.)
A word is said to be "key" if
a) it occurs in the text at least as many times as the user has specified as a Minimum Frequency
b) its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure 245 is smaller than or equal to a p value 235 specified by the user.

positive and negative keyness

A word which is *positively* key occurs *more* often than would be expected by chance in comparison with the reference corpus.
A word which is *negatively* key occurs *less* often than would be expected by chance in comparison with the reference corpus.

typical key words

KeyWords will usually throw up 3 kinds of words as "key".
First, there will be proper nouns. Proper nouns are often key in texts, though a text about racing could wrongly identify as key, names of horses which are quite incidental to the story. This can be avoided by specifying a higher Minimum Frequency.

Second, there are key words that human beings would recognise. The program is quite good at finding these, and they give a good indication of the text's "aboutness". (All the same, the program does not group synonyms, and a word which only occurs once in a text may sometimes be "key" for a human being. And **KeyWords** will not identify key phrases unless you are comparing word-lists based on word clusters 448.)

Third, there are high-frequency words like **because** or **shall** or **already**. These would not usually be identified by the reader as key. They may be key indicators more of style than of "aboutness". But the fact that KeyWords identifies such words should prompt you to go back to the text, perhaps with Concord (just choose *Compute* | *Concordance* C), to investigate why such words have cropped up with unusual frequencies.

See also: How Key Words are Calculated 245, Definition of Key Key-Word 240, Definitions 425, KeyWords Settings 254

8.7.3 thinking about keyness

Choosing a reference corpus

In general the choice does not make a lot of difference if you have a fairly small p value (such as 0.000001). But it may help to think using this analogy.

Different reference corpora may give different results. Suppose you have a method for comparing objects and you take a particular apple out of your kitchen to compare using it

A) with a lot of apples in the green-grocer's shop

B) with all the fruit in the green-grocer's shop

C) with a mixture of objects (cars, carpet, notebooks, fruit, elephants etc.)

With A) you will get to see the individual characteristics, e.g. perhaps your apple is rather sweeter than most apples. (But you won't see its "apple-ness" because both your apple and all the others in your reference corpus are all apples.)

With B) you will see "appleness" (your apple, like all apples but unlike bananas or pineapples, is rather round and has a very thin skin) but might not see that your apple is rather sweet and you won't get at its "fruitiness".

With C) you will get at the apple's fruity qualities: it is much sweeter and easier to bite into than cars and notebooks etc.

Keyness scores

Is there an important difference between a key word with a keyness of 50 and another of 500?

Suppose you process a text about a farmer growing 3 crops (wheat, oats and chick-peas) and suffering from 3 problems (rain, wind, drought). If each of these crops is equally important in the text, and each of the 3 problems takes one paragraph each to explain, the human reader may decide that all three crops are equally key and all three problems equally key. But in English these three crop-terms and weather-terms vary enormously in frequency (chick-peas and drought least frequent). WordSmith's KW analysis will necessarily give a higher keyness value to the rarer words. So it is generally unsafe to rely on the order of KWs in a KW list.

8.8 KeyWords database

(default file extension .KDB)

The point of it...

The point of this database is that it will allow you to study the key-words which recur often over a number of files.

For example, if you have 500 business reports, each one will have its own key words. These will probably be of two main kinds. There will be key-words which are key in one text but are not generally key (names of the firms and words relating to what they individually produce); and other, more general words (like **consultant, profit, employee**) which are typical of business documentation generally. Or you may find that **I, you, should** etc. come to the top if your text files are ones which are much more interactive than the reference corpus texts.

By making up a database, you can sort these out. The ones at the top of the list, when you view them, may be those which are most typical of the genre in some way. We might call the ones at the top "key-key words" and the list is at first ordered in terms of "key key-ness", but those at the bottom will only be key in a few text files. You can of course toggle it into alphabetical order and back again.

You can set a minimum number of files that each word must have been found to be key in, using *KeyWords Settings* | *Database* 238.

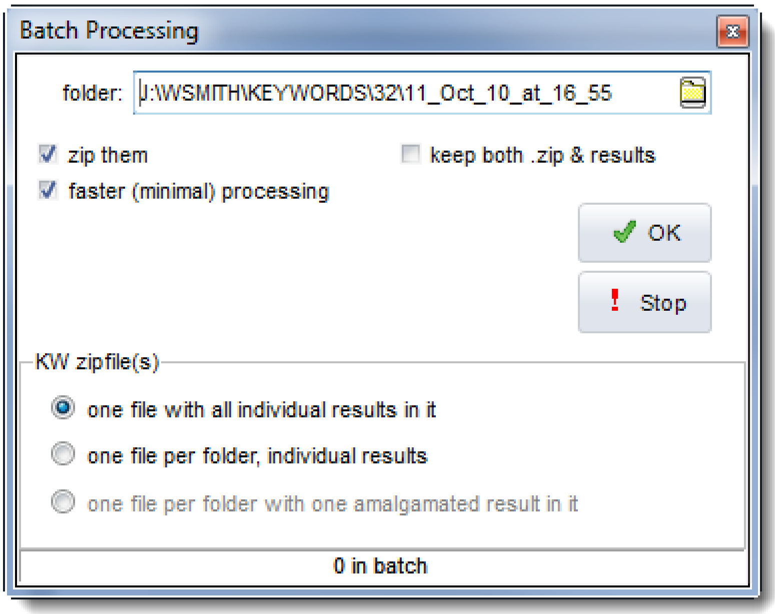
When viewing a database you will be able to investigate the associates 241 of the key key-words. Under Statistics, you will also be able to see details of the key words files which comprise the database (file name and number of key words per file), together with overall statistics on the number of different types and the tokens (the total of all the key-words in the whole database including repeats).

See also : Creating a database 238, Definition of key key-word 240

8.8.1 creating a database

To build a key words database, you will need a set of key word lists. For a decent sized database, it is preferable to build it like this:

1. Make a batch 39 of word lists.
2. Use this to make a batch 39 of keyword lists. Set "faster minimal processing" on as in this shot, so as to not waste time computing plots etc.



Batch Processing

folder: J:\WSMITH\KEYWORDS\32\11_Oct_10_at_16_55

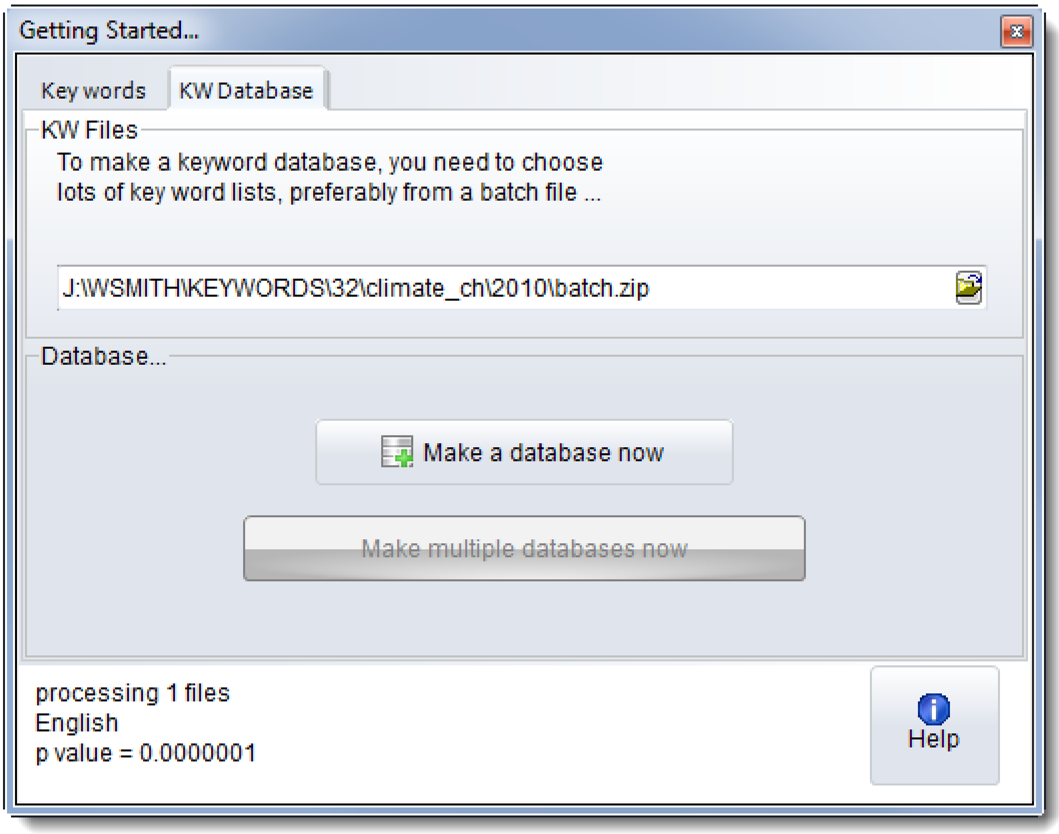
☑ zip them
☐ keep both .zip & results
☑ faster (minimal) processing

KW zipfile(s)

● one file with all individual results in it
○ one file per folder, individual results
○ one file per folder with one amalgamated result in it

0 in batch

3. Now, in **KeyWords**, choose *New* | *KW Database*.



To make a keyword database, you need to choose lots of key word lists, preferably from a batch file ...

J:\WSMITH\KEYWORDS\32\climate_ch\2010\batch.zip

processing 1 files
English
p value = 0.0000001

This enables you to choose the whole set of key word files.
Note that making a database means that only positive 245 key words will be retained.

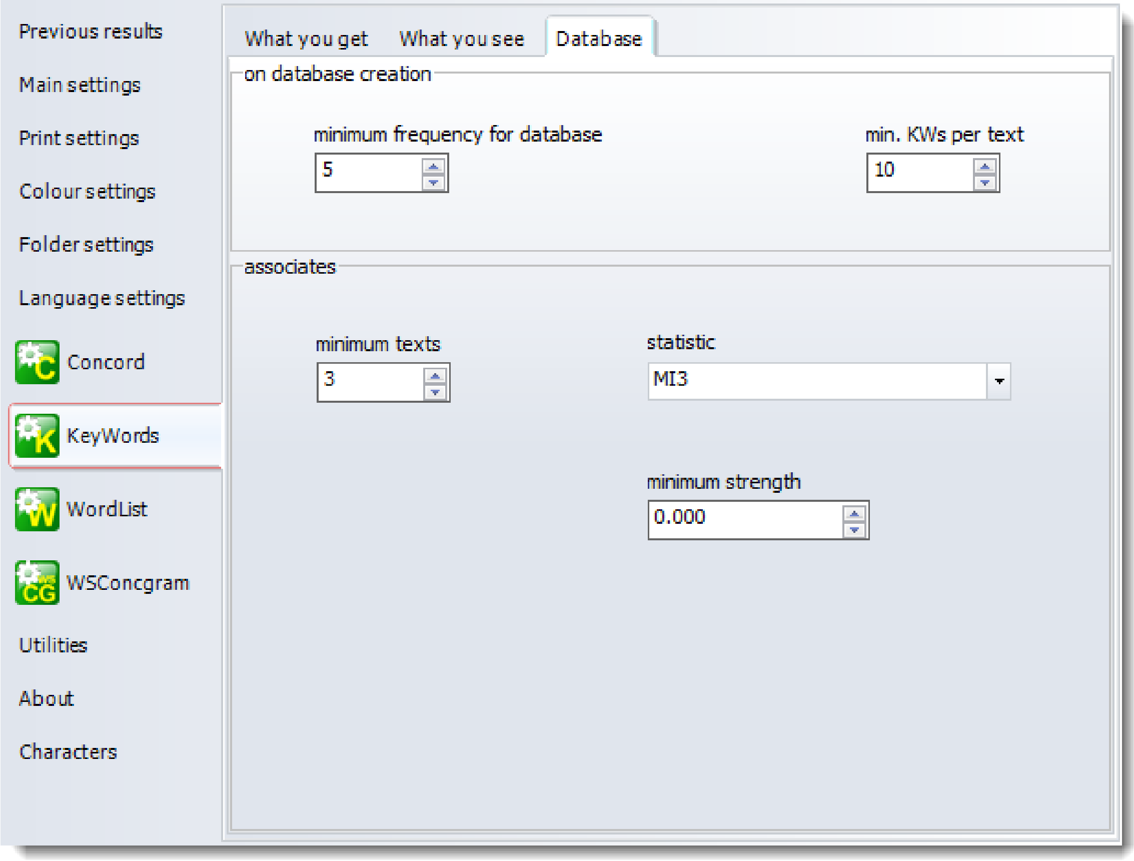
In the Controller KeyWords settings 254 you can make other choices:

minimum frequency for database

If you set this to 5 you will only use for the database any KWs which appear in 5 or more texts

min. KWs per text

If this is set to 10, any KW results files which ended up with very few positive KWs will be ignored.



on database creation

minimum frequency for database: 5

min. KWs per text: 10

associates

minimum texts: 3

statistic: MI3

minimum strength: 0.000

See also: associates 241.

8.8.2 key key-word definition

A "key key-word" is one which is "key" in more than one of a number of related texts. The more texts it is "key" in, the more "key key" it is. This will depend a lot on the topic homogeneity of the corpus being investigated. In a corpus of City news texts, items like *bank*, *profit*, *companies* are key key-words, while *computer* will not be, though *computer* might be a key word in a few City news stories about IBM or Microsoft share dealings.

Requirements

To discover "key key words" you need a lot of text files (say 500 or more), ideally fairly related in their topics, which you make word-lists of (it's much faster doing that in a batch), and then you have to compute key word-lists of each of those, all of which go into a database. It is all explained under creating a keywords database 238.

See also: How Key Words are Calculated 245, Definition of Key Word 236, Creating a Database 238, Definitions 425

8.8.3 associates

"Associates" is the name given to key-words associated with a key key-word 237.

The point of it...

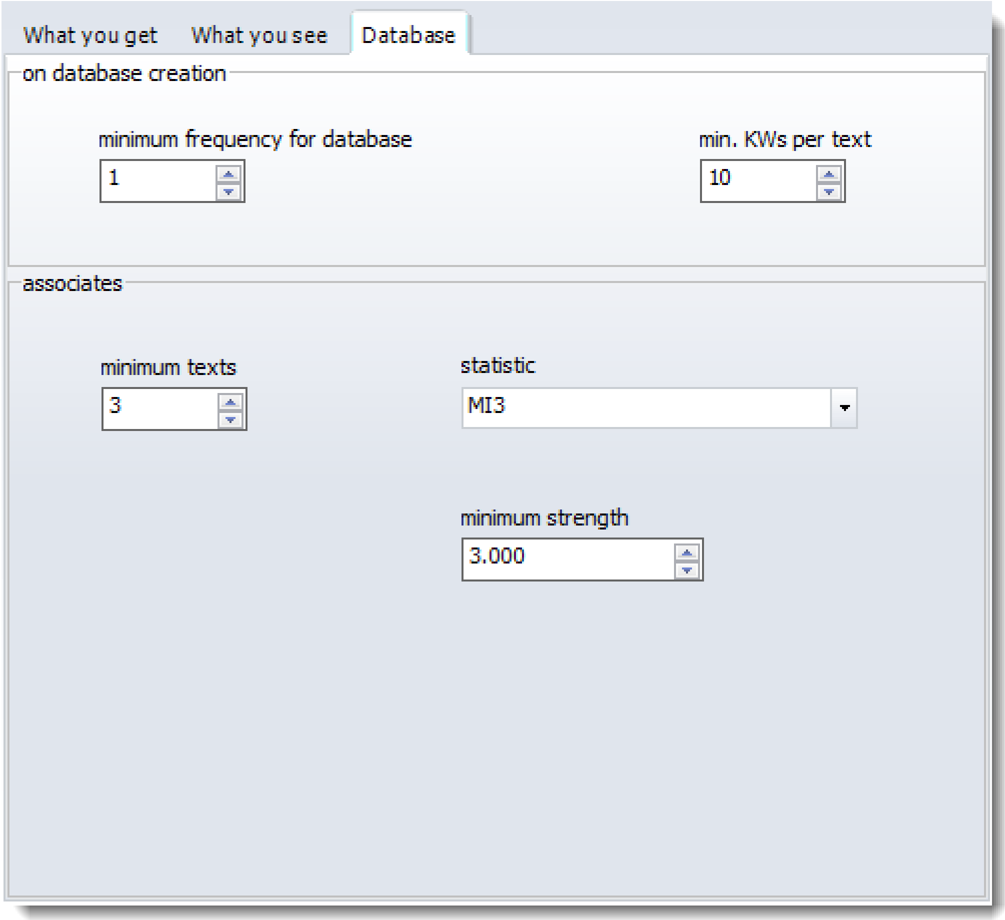
The idea is to identify words which are commonly associated with a key key-word, because they are key words in the same texts as the key key-word is. An example will help.

Suppose the word *wine* is a key key-word in a set of texts, such as the weekend sections of newspaper articles. Some of these articles discuss different wines and their flavours, others concern cooking and refer to using wine in stews or sauces, others discuss the prices of wine in a context of agriculture and diseases affecting vineyards. In this case, the associates of *wine* would be items like *Chardonnay, Chile, sauce, fruit, infected, soil*, etc.

The listing shows associates in order of frequency. A menu option allows you to re-sort them.

Settings

You can set a minimum number of text files for the association procedure, in the database settings 254:



on database creation

minimum frequency for database: 1

min. KWs per text: 10

associates

minimum texts: 3

statistic: MI3

minimum strength: 3.000

Minimum texts

These screenshot settings would only process those key-key-words which appear in at least 3 text files.

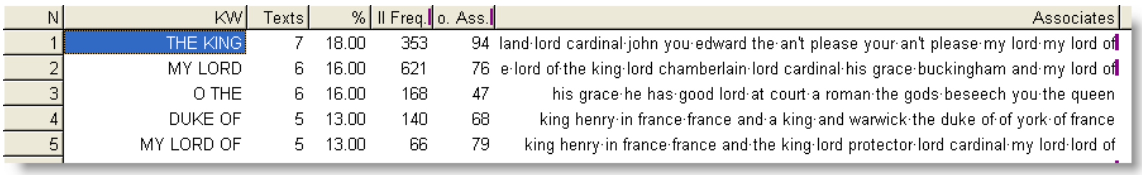
Statistic

Choose the mutual information statistic 289 you prefer, apart from Z score which uses a span (here we're using the whole text).

Minimum strength

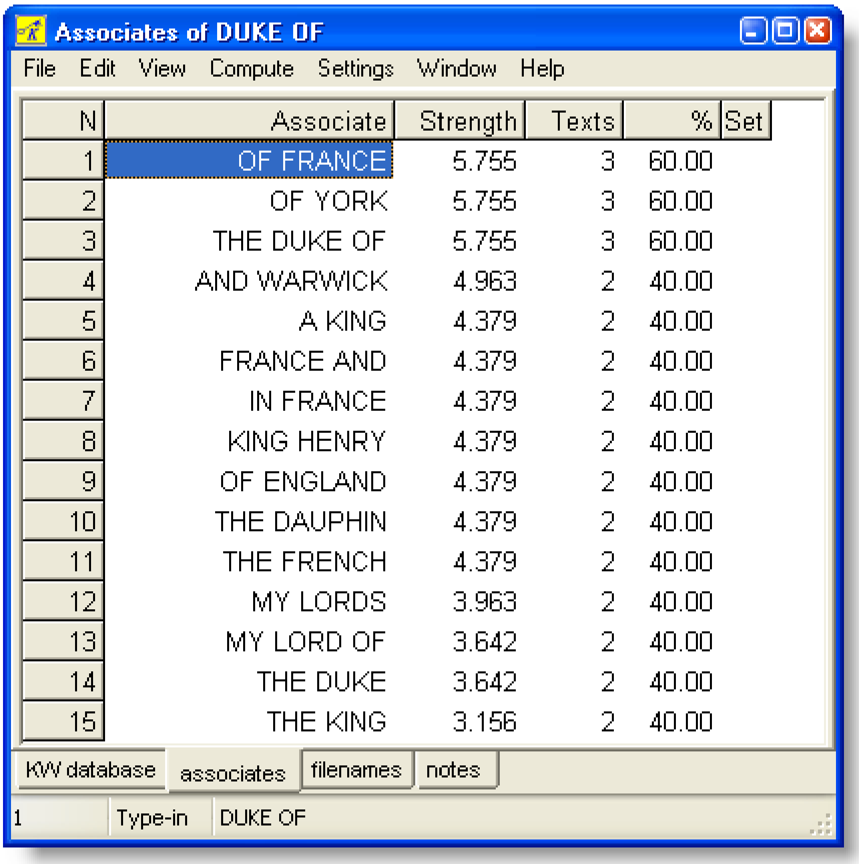
This will only show associates which reach at least the strength in the statistic set here, e.g. 3.000.

This screenshot shows the most frequent associates in the right-hand column of the main keywords data base window.



N	KW	Texts	%	ll Freq.	o. Ass.	Associates
1	THE KING	7	18.00	353	94	land·lord cardinal·john you·edward the·an't please your·an't please·my lord·my lord of
2	MY LORD	6	16.00	621	76	e·lord of the king·lord chamberlain·lord cardinal·his grace·buckingham and·my lord of
3	O THE	6	16.00	168	47	his grace·he has·good lord·at court·a roman·the gods·beseech you·the queen
4	DUKE OF	5	13.00	140	68	king henry·in france·france and·a king·and warwick·the duke of·of york·of france
5	MY LORD OF	5	13.00	66	79	king henry·in france·france and·the king·lord protector·lord cardinal·my lord·lord of

To see the detailed associates, double-click your chosen term in the KW column or the Associates column:



N	Associate	Strength	Texts	%	Set
1	OF FRANCE	5.755	3	60.00	
2	OF YORK	5.755	3	60.00	
3	THE DUKE OF	5.755	3	60.00	
4	AND WARWICK	4.963	2	40.00	
5	A KING	4.379	2	40.00	
6	FRANCE AND	4.379	2	40.00	
7	IN FRANCE	4.379	2	40.00	
8	KING HENRY	4.379	2	40.00	
9	OF ENGLAND	4.379	2	40.00	
10	THE DAUPHIN	4.379	2	40.00	
11	THE FRENCH	4.379	2	40.00	
12	MY LORDS	3.963	2	40.00	
13	MY LORD OF	3.642	2	40.00	
14	THE DUKE	3.642	2	40.00	
15	THE KING	3.156	2	40.00	

1 Type-in DUKE OF

See also: definition of associate 243, related clusters 243.

8.8.3.1 associate definition

An "associate" of key-word X is another key-word (Y) which co-occurs with X in a number of texts. It may or may not co-occur in proximity to key-word X. (A *collocate* would have to occur within a given distance of it, whereas an associate is "associated" by being key in the same text.)

For example, in a key-word database of *Guardian* newspaper text, *wine* was found to be a key word in 25 out of 299 stories from the Saturday "tabloid" page, thus a key key word 240 in this section. The top associates of *wine* were: *wines, Tim, Atkin, dry, le, bottle, de, fruit, region, chardonnay, red, producers, beaujolais*.

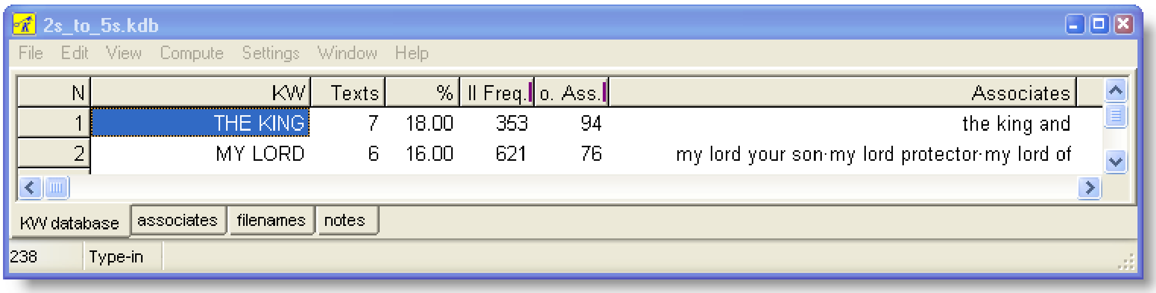
It is strikingly close to the early notion of "collocate".

Association operates in various ways. It can be strong or weak, and it can be one-way or two-way. For example, the association between *to* and *fro* is one-way (*to* is nearly always found near *fro* but it is rare to find *fro* near *to*).

See also: Definition of Key Word 236, Associates 241, Definitions 425, Mutual Information 289

8.8.4 keywords database related clusters

The idea is to be able to find any overlapping clusters in a key word database, e.g. where MY LORD is related to MY LORD YOUR SON.



N	KW	Texts	%	ll Freq.	o. Ass.	Associates
1	THE KING	7	18.00	353	94	the king and
2	MY LORD	6	16.00	621	76	my lord your son·my lord protector·my lord of

KW database | associates | filenames | notes

238 Type-in

To achieve this, choose *Compute* | *Clusters*. To clear the view, *Compute* | *Associates*.

See also: associates 241

8.8.5 clumps

"Clumps" is the name given to groups of key-words associated 241 with a key key-word 237.

The point of it (1)...

The idea here is to refine associates by grouping together words which are found as key in the same sub-sets of text files. The example used to explain associates will help.

Suppose the word *wine* is a key key-word in a set of texts, such as the weekend sections of newspaper articles. Some of these articles discuss different wines and their flavours, others concern cooking and refer to using wine in stews or sauces, others discuss the prices of wine in a context of agriculture and diseases affecting vineyards. In this case, the associates of *wine* would be items like *Chardonnay, Chile, sauce, fruit, infected, soil*, etc. The associates procedure shows all such items unsorted.

The clumping procedure, on the other hand, attempts to sort them out according to these different uses. The reasoning is that the key words of each text file give a condensed picture of its "aboutness", and that "aboutnesses" of different texts can be grouped by matching the key word lists. Thus sets of key words can be clumped together according to the degree of overlap in the key word lexis of each text file.

Two stages

The **initial clumping process does no grouping**: you will simply see each set of key-words for each text file separately. To group clumps 244, you may simply join those you think belong together (by dragging), or regroup with help by pressing .
The listing shows clumps sorted in alphabetical order. You can re-sort by frequency (the number of times each key word in the clump appeared in all the files which comprise the clump).
See also: definition of associate 243, regrouping clumps 244

8.8.5.1 regrouping clumps

How to do it

You can simply join by dragging, where you think any two clumps belong together because of semantic similarity between their key-words.
Or if you press , **KeyWords** will inform you which two clumps match best. You'll see a list of the words found only in one, a list of the words found only in the other, and (in the middle) a list of the words which match. It's up to you to judge whether the match is good enough to form a merged clump.
If you aren't sure, press **Cancel**.
If you do want to join them, press **Join**.
If you're sure you **don't** want to join them and don't want **KeyWords** to suggest this pair again, press **Skip**. You can tell **KeyWords** to skip up to 50 pairs. To clear the memory of the items to be skipped, press **Clear Skip**.

The point of it (2)...

Scott 417 (1997) shows how clumping reveals the different perceived roles of women in a set of *Guardian* features articles.
See also: clumps 243

8.9 KeyWords: advice

1. Don't call up a plot of the key words based on more than one text file. It doesn't make sense! Anyway the plot will only show the words in the first text file. If you want to see a plot of a certain word or phrase in various different files, use Concord dispersion 191.

2. There can be no guarantee that the "key" words are "key" in the sense which you may attach to "key". An "important" word might occur once only in a text. They are merely the words which are outstandingly frequent or infrequent in comparison with the reference corpus.

3. Compare apples with pears, or, better still, Coxes with Granny Smiths. So choose your reference corpus in some principled way 237. The computer is not intelligent and will try to do whatever comparisons you ask it to, so it's up to you to use human intelligence and avoid comparing apples with phone boxes!

If it didn't work...

For the procedure to work, a number of conditions must be right: the language 81 defined for each word list must be the same (that is, Mexican Spanish and Iberian Spanish count as the same but Iberian Spanish and Brazilian Portuguese count as different so could not be compared in this process); each word list must have been sorted alphabetically 315 in ascending order before the comparison is made. (The program tries to ensure this, automatically.) Also, any prefixes 315 or suffixes must match.

8.10 KeyWords: calculation

The "key words" are calculated by comparing the frequency of each word in the word-list of the text you're interested in with the frequency of the same word in the reference word-list. All words which appear in the smaller list are considered, unless they are in a stop list 120.

If **the** occurs say, 5% of the time in the small word-list and 6% of the time in the reference corpus, it will not turn out to be "key", though it may well be the most frequent word. If the text concerns the anatomy of spiders, it may well turn out that the names of the researchers, and the items **spider, leg, eight**, etc. may be more frequent than they would otherwise be in your reference corpus (unless your reference corpus only concerns spiders!)

To compute the "key-ness" of an item, the program therefore computes

its frequency in the small word-list
the number of running words 297 in the small word-list
its frequency in the reference corpus
the number of running words 297 in the reference corpus

and cross-tabulates these.

Statistical tests include:

the classic chi-square test of significance with Yates correction for a 2 X 2 table
Ted Dunning's 417 Log Likelihood test, which gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus.

See UCREL's log likelihood site for more on these.

A word will get into the listing here if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger word-list.

Unusually infrequent key-words are called "negative key-words" and appear at the very end of your listing, in a different colour. Note that negative key-words will be omitted automatically from a keywords database 237 and a plot.

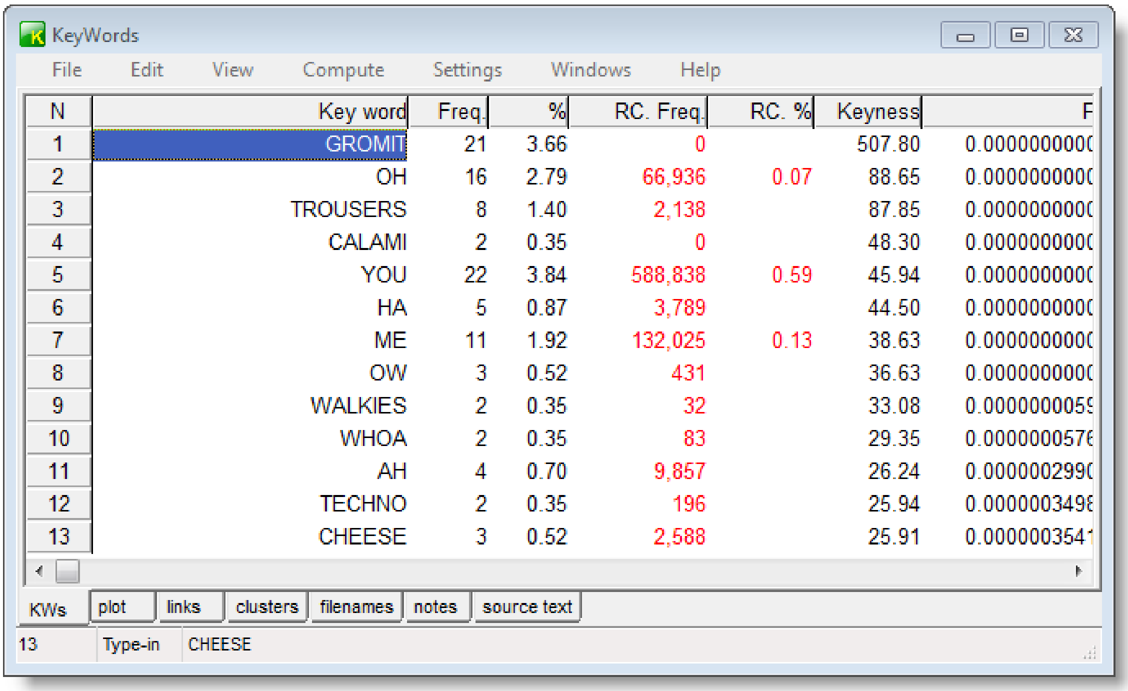
Words which do not occur at all in the reference corpus are treated as if they occurred 5.0e-324 times (0.0000000 and loads more zeroes before a 5) in such a case. This number is so small as not to affect the calculation materially while not crashing the computer's processor.

8.11 KeyWords clusters

What is it?

A KeyWords cluster, like a WordList cluster, represents two or more words which are found repeatedly near each other. However, a **KeyWords** cluster only uses key words.

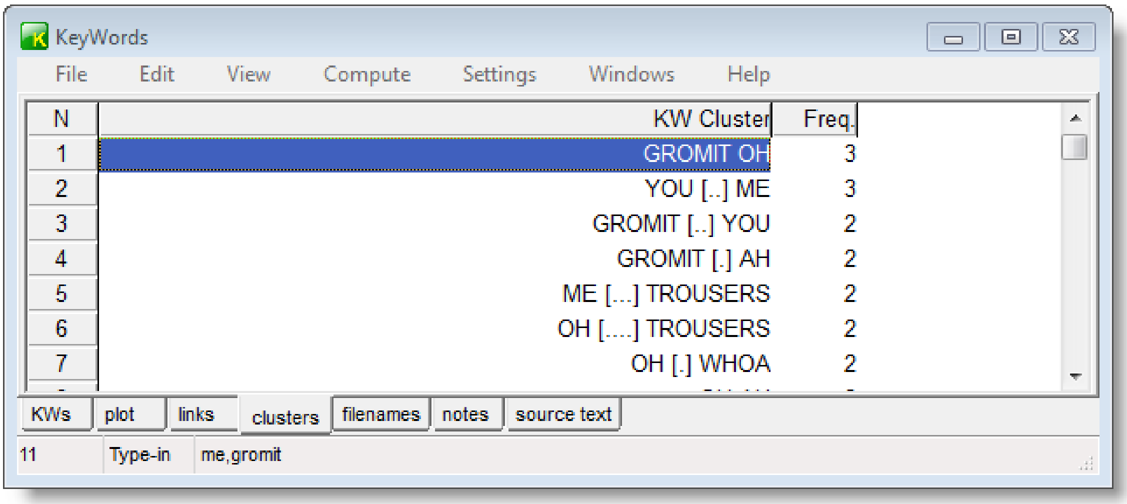
A screenshot will help make things clearer. This is a key words list based on a piece of transcript from a Wallace and Gromit film, using the BNC as the reference corpus.



N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	P
1	GROMIT	21	3.66	0		507.80	0.0000000000
2	OH	16	2.79	66,936	0.07	88.65	0.0000000000
3	TROUSERS	8	1.40	2,138		87.85	0.0000000000
4	CALAMI	2	0.35	0		48.30	0.0000000000
5	YOU	22	3.84	588,838	0.59	45.94	0.0000000000
6	HA	5	0.87	3,789		44.50	0.0000000000
7	ME	11	1.92	132,025	0.13	38.63	0.0000000000
8	OW	3	0.52	431		36.63	0.0000000000
9	WALKIES	2	0.35	32		33.08	0.0000000059
10	WHOA	2	0.35	83		29.35	0.0000000576
11	AH	4	0.70	9,857		26.24	0.0000002990
12	TECHNO	2	0.35	196		25.94	0.0000003498
13	CHEESE	3	0.52	2,588		25.91	0.0000003541

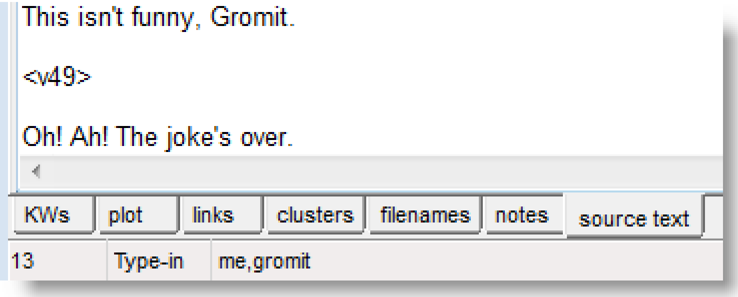
13 Type-in CHEESE

The clusters tab below shows us something like this:



N	KW Cluster	Freq.
1	GROMIT OH	3
2	YOU [..] ME	3
3	GROMIT [..] YOU	2
4	GROMIT [.] AH	2
5	ME [...] TROUSERS	2
6	OH [....] TROUSERS	2
7	OH [.] WHOA	2

The frequency 3 in the **GROMIT OH** line means that there are 3 cases where the key-word **GROMIT** is found within the current collocation span of **OH** in that text. **[.]** means that there is typically one intervening word or [..] two intervening words as in this case shown from the source text.


This isn't funny, Gromit.

<v49>

Oh! Ah! The joke's over.

Requirements

The procedure is text-oriented. You can only get a keywords cluster list if there is exactly one source text. Note that for this procedure sentence boundaries are not blocked, so **Gromit** and **Ah** can be considered to have one word **Oh** intervening.

See also: Plot calculation 251.

8.12 KeyWords: links

The point of it...

is to find out which key-words are most closely related to a given key-word.

A plot 251 will show where each key word occurs in the original file. It also shows how many links there are between key-words.

What are links?

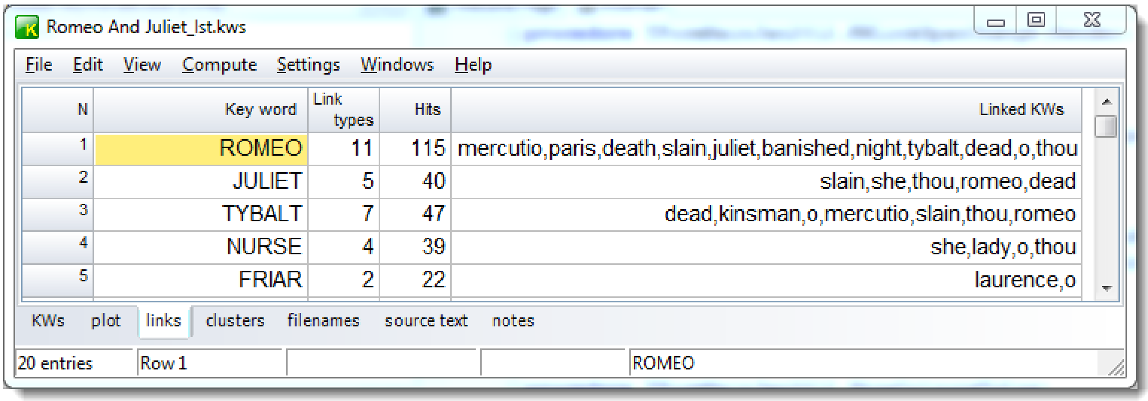
Links are "co-occurrences of key-words within a collocational span". An example is much easier to understand, though:

Suppose the word *elephant* is key in a text about Africa, and that *water* is also a key word in the same text. If *elephant* and *water* occur within a span of 5 words of each other, they are said to be "linked". The number of times they are linked like this in the text will be shown in the Links window.

The link spans (like collocation horizons) go from 1 word away to up to 25 words to left and right. The default 113 is 1 to 5.

What you see

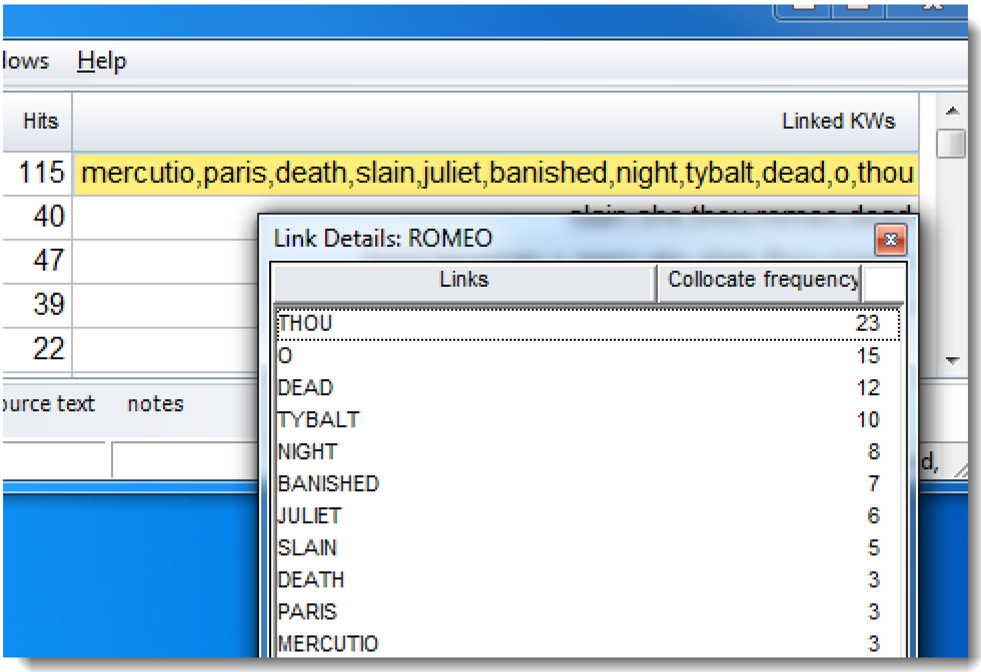
This is a key words list based on Romeo and Juliet, using all the 37 Shakespeare plays as the reference corpus.


Romeo And Juliet_lst.kws
File Edit View Compute Settings Windows Help
KWs plot links clusters filenames source text notes

N	Key word	Link types	Hits	Linked KWs
1	ROMEO	11	115	mercutio,paris,death,slain,juliet,banished,night,tybalt,dead,o,thou
2	JULIET	5	40	slain,she,thou,romeo,dead
3	TYBALT	7	47	dead,kinsman,o,mercutio,slain,thou,romeo
4	NURSE	4	39	she,lady,o,thou
5	FRIAR	2	22	laurence,o

20 entries Row 1 ROMEO

This Links window shows a number of key words followed by the number of linked types (11 here) the total number of hits of the key word (**ROMEO**) and then the individual linked key words. You can if you wish double-click in the *Linked KWs* column and you will see the details listed:


lows Help
Hits
Linked KWs
115 mercutio,paris,death,slain,juliet,banished,night,tybalt,dead,o,thou
40
47
39
22
urce text notes

Link Details: ROMEO

Links	Collocate frequency
THOU	23
O	15
DEAD	12
TYBALT	10
NIGHT	8
BANISHED	7
JULIET	6
SLAIN	5
DEATH	3
PARIS	3
MERCUTIO	3

ROMEO has 11 linked words; it's linked 23 times with **THOU**, 15 times with **O**, etc. A right-click menu lets you copy or print these details.

Requirements

The procedure is text-oriented. You can only get a keywords links list if there is exactly one source text.

Double-click on any word in the plot listing[251] to call up a window which show the linked key-words.

See also: Plot calculation[251], KeyWords clusters[246], Source Text[116]

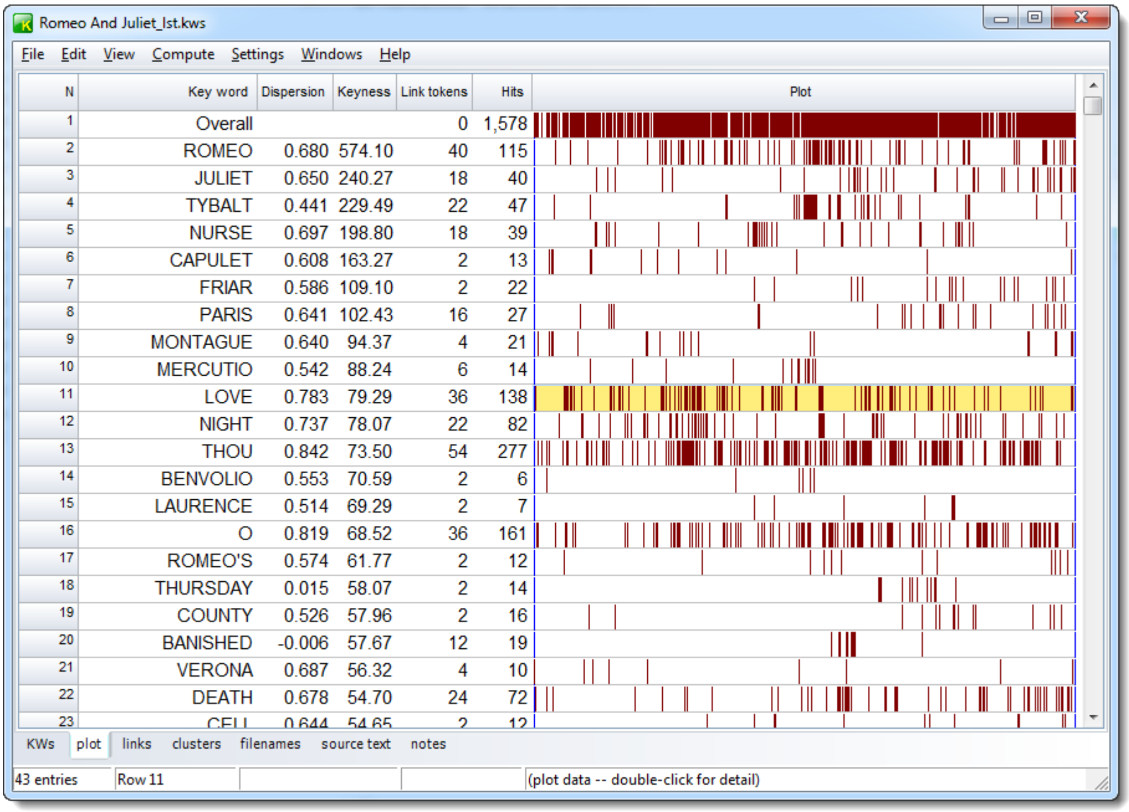
8.13 make a word list from keywords data

With a key word list on your screen, you can press W to save your data as a word list (for later comparison, etc. using **WordList** functions).

8.14 plot display

The plot will give you useful visual insights into how often and where the different key words crop up in the text. The plot is initially sorted[252] to show which crop up more at the beginning (e.g. in the introduction) and then those from further in the text.

The following screenshot shows KWs of the play *Romeo and Juliet*, revealing where each term occurs. The name **Tybalt**, for example, occurs in a main burst about half way through the text.



Romeo And Juliet_lst.kws

N	Key word	Dispersion	Keyness	Link tokens	Hits	Plot
1	Overall			0	1,578	
2	ROMEO	0.680	574.10	40	115	
3	JULIET	0.650	240.27	18	40	
4	TYBALT	0.441	229.49	22	47	
5	NURSE	0.697	198.80	18	39	
6	CAPULET	0.608	163.27	2	13	
7	FRIAR	0.586	109.10	2	22	
8	PARIS	0.641	102.43	16	27	
9	MONTAGUE	0.640	94.37	4	21	
10	MERCUTIO	0.542	88.24	6	14	
11	LOVE	0.783	79.29	36	138	
12	NIGHT	0.737	78.07	22	82	
13	THOU	0.842	73.50	54	277	
14	BENVOLIO	0.553	70.59	2	6	
15	LAURENCE	0.514	69.29	2	7	
16	O	0.819	68.52	36	161	
17	ROMEO'S	0.574	61.77	2	12	
18	THURSDAY	0.015	58.07	2	14	
19	COUNTY	0.526	57.96	2	16	
20	BANISHED	-0.006	57.67	12	19	
21	VERONA	0.687	56.32	4	10	
22	DEATH	0.678	54.70	24	72	
23	CELL	0.644	54.65	2	12	

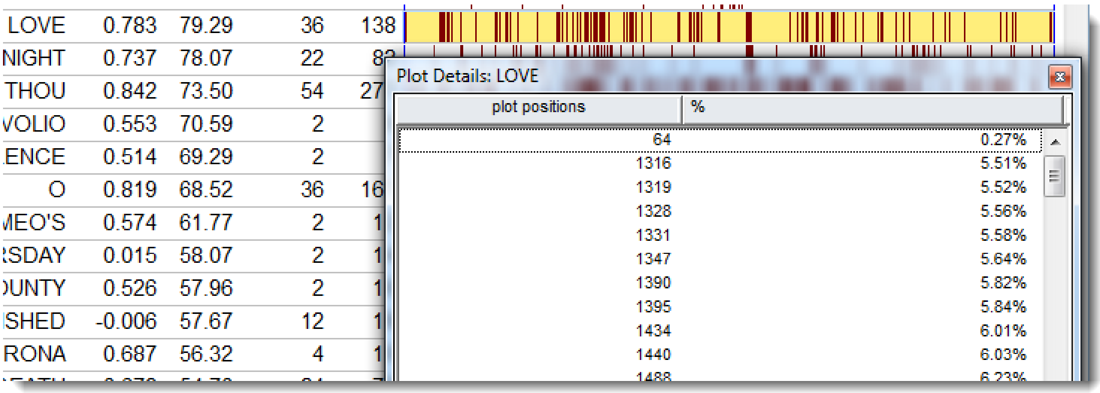
43 entries | Row 11 | (plot data -- double-click for detail)

re-sorting

Click the header to re-sort 252 the listing or use the menu **option**. The Key word column sorts alphabetically, the dispersion column sorts on the amount of dispersion (higher numbers mean the occurrences are more spread out); the keyness column is the original plot order, or you can sort on number of links with other KWs or on the number of hits found.

plot data

You can view the plot data as numbers by double-clicking. Here is the view if one double-clicks on the yellow area:



LOVE	0.783	79.29	36	138
NIGHT	0.737	78.07	22	8
THOU	0.842	73.50	54	27
VOLIO	0.553	70.59	2	
ENCE	0.514	69.29	2	
O	0.819	68.52	36	16
MEO'S	0.574	61.77	2	1
RSDAY	0.015	58.07	2	1
OUNTY	0.526	57.96	2	1
SHED	-0.006	57.67	12	1
RONA	0.687	56.32	4	1

Plot Details: LOVE

plot positions	%
64	0.27%
1316	5.51%
1319	5.52%
1328	5.56%
1331	5.58%
1347	5.64%
1390	5.82%
1395	5.84%
1434	6.01%
1440	6.03%
1488	6.23%

The first column gives the word-numbers and the second the percentage of the way through the text. Right-click on this window to copy or print.

links

This shows the total number of links 247 between the key-word and other key-words in the same text, within the current collocation span (default 113 = 5,5). That is, how many times was each key-word found within 5 words of left or right of any of the other key-words in your plot.

hits

This column is here to remind you of how many occurrences there were of each key-word.

When you have obtained a plot, you can then see the way certain words relate to others. To do this, look at the Links window in the tabs at the bottom, showing which other key words are most linked 247 to the word you clicked on. That is, which other words occur most often within the collocation horizons you've set. The Links window should help you gain insights into the lexical relations here.

Each plot window is dependent on the key words listing from which it was derived. If you close that down, it will disappear. You can *Print* it. There's no *Save* option because the plot comes from a key words listing which you should *Save*, or *Save As*. There's no save as text 102 option because the plot has graphics, which cannot adequately be represented as text symbols, but you can *Copy* to the clipboard 422 (Ctrl+C) and then paste it into a word processor as a graphic. Alternatively, use the *Output* | *Data as Text File* option, which saves your plot data (each word is followed by the total number of words in the file, then the word number position of each occurrence).

The ruler 441 in the menu () allows you to see the plot divided into 8 equal segments if based on one text, or the text-file divisions if there is more than one.

See also: Key words plot 251, plot dispersion value 446

8.14.1 plot calculation

The point of it...

is to see where the key words are distributed within the text. Do they cluster around the middle or near the beginning of the text?

How it's done

This will calculate the inter-relationships between all the key words identified so far, excluding any which you have deleted or zapped 129.

1. it does a concordance on the text finding all occurrences of each key word;
2. it then works out which of each of the other key words appear within the collocation horizons (set in Settings). It uses the larger of the two horizons.
3. it then plots all the words showing where each occurrence comes in the original file (with a "ruler" showing how many words there are in each part of the file).
4. it computes how many other key-words co-occurred with it, within the current collocational span.
5. it computes a plot dispersion value 446.

Note: this process depends on KeyWords being able to find the source texts 430 which your original word-list was based on.

You may find it useful to export your plot 102 and make other graphs, as explained under Save As 102.

See also: Plot Links 247, Key words plot display 249

8.15 re-sorting: KeyWords

How to do it...

Sorting can be done simply by pressing the top row of any list. Or by pressing F6 or Ctrl+F6. Or by choosing the menu option. Press again to toggle between ascending & descending sorts.

the different sorts

A **key words list** offers a choice between sorting by

key-ness (the *key*est words appear at the top)
alphabetical order (from A to Z)
frequency in the smaller list (the most frequent words come first)
frequency in the reference list (the most frequent words come first)

A **key words plot** rotates between sorting by

key-ness (the *key*est words appear at the top)
alphabetical order (from A to Z)
frequency (words which appear oftenest come first)
number of links (the most linked words come first)
first mention of each key word in the text
range (words used in smallest sections of text come first)

A **key key words database** toggles between sorting by

frequency (the most *key key* words appear at the top)
alphabetical order (from A to Z)

An **Associates** 241 **list** toggles between sorting by

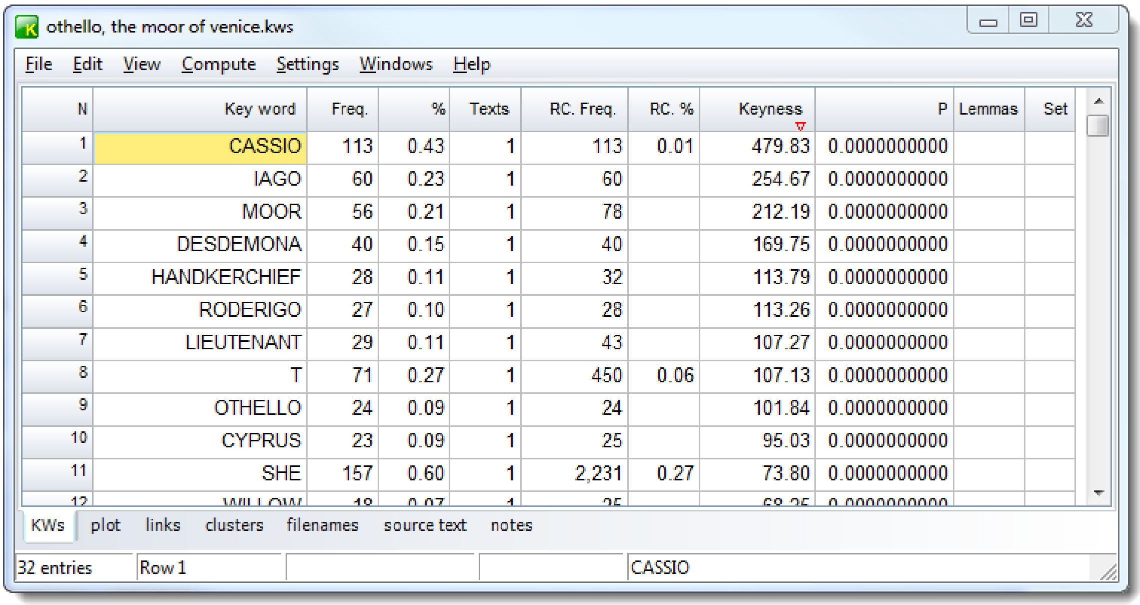
frequency (association between title-word and item)
alphabetical order (from A to Z)
frequency (association between item and title-word)

8.16 the key words screen

The display shows

- each key word
- its frequency in the source text(s) which these key words are key in. (Freq. column below)
- the % that frequency represents
- the number of texts it was present in
- its frequency in the reference corpus (RC. Freq. column)
- the reference corpus frequency as a %

- keyness (chi-square or log likelihood statistic 245)
- p value 235
- lemmas 270 (any which have been joined to each other)
- the user-defined set 168



N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	Keyness	P	Lemmas	Set
1	CASSIO	113	0.43	1	113	0.01	479.83	0.0000000000		
2	IAGO	60	0.23	1	60		254.67	0.0000000000		
3	MOOR	56	0.21	1	78		212.19	0.0000000000		
4	DESDEMONA	40	0.15	1	40		169.75	0.0000000000		
5	HANDKERCHIEF	28	0.11	1	32		113.79	0.0000000000		
6	RODERIGO	27	0.10	1	28		113.26	0.0000000000		
7	LIEUTENANT	29	0.11	1	43		107.27	0.0000000000		
8	T	71	0.27	1	450	0.06	107.13	0.0000000000		
9	OTHELLO	24	0.09	1	24		101.84	0.0000000000		
10	CYPRUS	23	0.09	1	25		95.03	0.0000000000		
11	SHE	157	0.60	1	2,231	0.27	73.80	0.0000000000		
12	WILLOW	18	0.07	1	25		68.25	0.0000000000		

32 entries | Row 1 | CASSIO

The calculation of how unusual the frequency is, is based on the statistical procedure 245 used. The statistic appears to the right of the display. If the procedure is log likelihood, or if chi-square is used and the usual conditions for chi-square obtain (expected value >= 5 in all four cells) the probability (p) will be displayed to the right of the chi-square value.

The criterion for what counts as "outstanding" is based on the minimum probability value selected before the key words were calculated. The smaller the number, the fewer key words in the display. Usually you'll not want more than about 40 key words to handle.

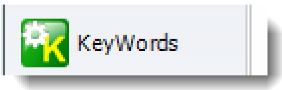
The words appear sorted 252 according to how outstanding their frequencies of occurrence are. Those near the top are outstandingly frequent. At the end of the listing you'll find any which are outstandingly infrequent 236 (negative keywords), in a different colour.

There is no upper limit to the keyness column of a set of key words. It is not necessarily sensible to assume that the word with the highest keyness value must be the most outstanding, since keyness is computed merely statistically; there will be cases where several items are obviously equally key (to the human reader) but the one which is found least often in the reference corpus and most often in the text itself will be at the top of the list.

Source text

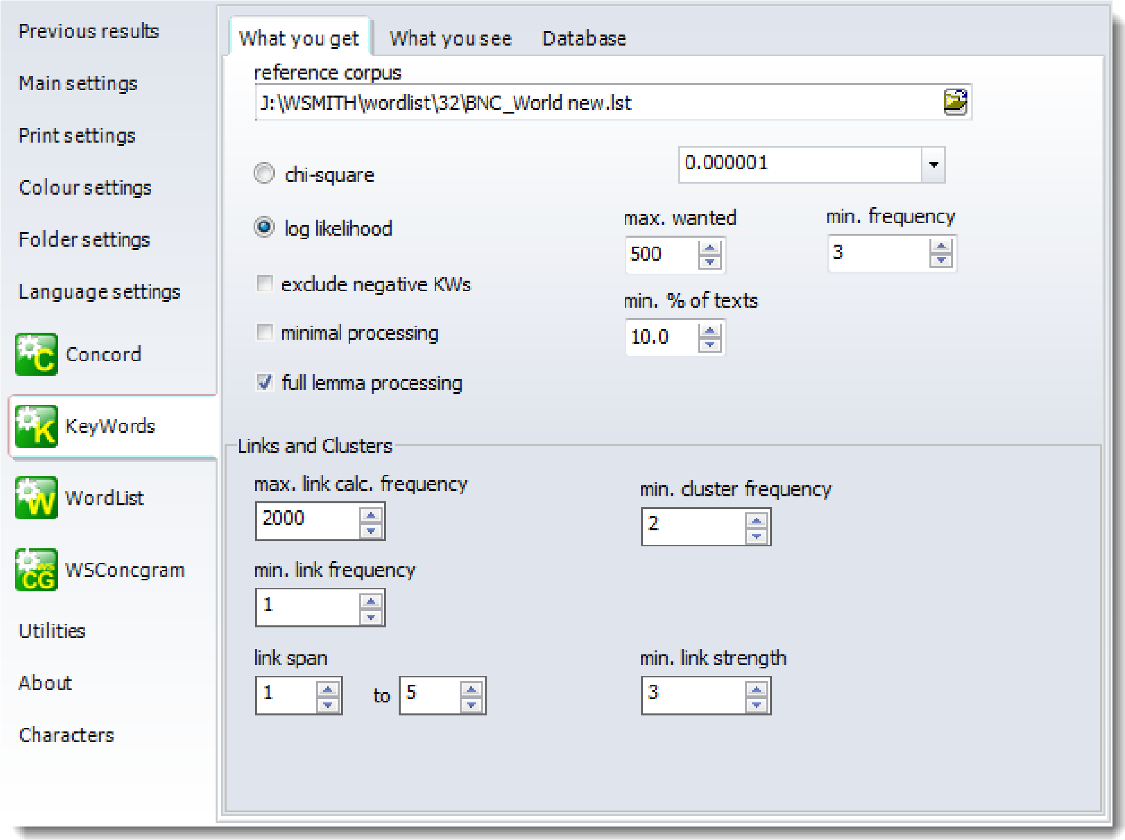
As its name suggests, choosing the source text tab gets you to a view of the source text 116(s).

8.17 WordSmith controller: KeyWords settings



These are found in the main Controller 4 marked *KeyWords.*

This is because some of the choices may affect other Tools. KeyWords and WordList both use similar routines: KeyWords to calculate the key words of a text file, and WordList when comparing word-lists 260.



reference corpus: J:\WSMITH\wordlist\32\BNC_World new.lst

○ chi-square

● log likelihood

0.000001

max. wanted: 500

min. frequency: 3

min. % of texts: 10.0

□ exclude negative KWs

□ minimal processing

☑ full lemma processing

Links and Clusters

max. link calc. frequency: 2000

min. cluster frequency: 2

min. link frequency: 1

link span: 1 to 5

min. link strength: 3

WHAT YOU GET

Procedure

Chi-square or Log Likelihood. The default is Log Likelihood. See procedure 245 for further details.

Max. p value

The default level of significance. See p value 235 for more details.

Max. wanted (500), Min. frequency (3), Min. % of texts (5%)

You may want to restrict the number of key words (KWs) identified so as to find for example the ten most "key" for each text. The program will identify all the key words, sort them by key-ness, and then throw away any excess. It will thus favour positive key words 236 over negative ones.

The minimum frequency is a setting which will help to eliminate any words or clusters which are unusual but infrequent. For example, a proper noun such as the name of a village will usually be extremely infrequent in your reference corpus, and if mentioned only once in the text you're analysing, it is likely not to be "key". The default setting of 3 mentions as a minimum helps reduce spurious hits here. In the case of short texts, less than 600 words long, a minimum of 2 will automatically be used.

The minimum percentage of texts (default = 5%) allows you to ignore words which are not found in many texts. Here the percentage is of the text files in the set you are comparing against a reference corpus. If you're comparing a word-list based on one text, each word in it will occur in 100% of the texts and thus won't get ignored. If you compare a word-list based on 200 texts against your reference corpus, the default of 5% would mean that only words which occur in at least 10 of those texts will be considered for keyness. The KeyWords display 252 shows the number of texts each KW was found in. (If you see ?? that is because the data were computed before that facility came into WordSmith.)

Exclude negative KWs

If this is checked, KeyWords will not compute negative key words (ones which occur significantly *in*frequently).

Minimal processing

If this is checked, KeyWords will not compute plots 251, links 247 or KW clusters 246 as it computes the key words (they can always be computed later assuming you do not move or delete the original text files). This is useful if computing a lot of KW files in a batch, eg. to make a database.

Full lemma processing

If this is checked (the default), KeyWords will compute the full frequency in the case of lemmatised 270 items. For example if **GO** represents **WENT, GOES** etc. and **GO** alone had a frequency of 10 but the whole set **GO, WENT, GONE** etc. totalled 100, then its frequency will be counted as 100. If unchecked, **GO** would count only 10.

Max. link frequency

To compute a plot is hard work as all the KWs have to be concordanced so as to work out where they crop up. To compute links between each KW is much harder work again and can take time especially if your KWs include some which occur thousands or hundreds of times in the text. To keep this process more manageable, you can set a default. Here 2000 means that any KW which occurs more than 2000 times in the text will not be used for computing links 247. (It will still appear in the plots and list of KWs, of course.)

WHAT YOU SEE

Columns

The *Columns to show/hide* list offers all the standard columns: you may uncheck ones you normally do not wish to see. This will only affect newly computed KeyWords data: earlier data uses the column visibility, size, colours etc already saved. They can be altered using the Layout 87 menu option at any time.

DATABASE

Database: minimum frequency

The default is 1. See database 237.

Database: associate minimum texts

The default is 5. See associates 241.

See also: KeyWords Help Contents 229, KeyWords calculation 245.

WordSmith Tools Manual

WordList

Section IX

9 WordList

9.1 purpose

This program generates word lists based on one or more plain text files. The word lists are automatically generated in both alphabetical and frequency order, and optionally you can generate a word index 276 list too.

The point of it...

These can be used

1. simply in order to study the type of vocabulary used;
2. to identify common word clusters 448;
3. to compare the frequency of a word in different text files or across genres;
4. to compare the frequencies of cognate words or translation equivalents between different languages 81;
5. to get a concordance 234 of one or more of the words in your list.

Within WordList you can compare two lists 260, or carry out consistency analysis (simple 262 or detailed 263) for stylistic comparison purposes.

These word-lists may also be used as input to the KeyWords 229 program, which analyses the words in a given text and compares frequencies with a reference corpus, in order to generate lists of "key-words" and "key-key-words".

Word lists don't have to be of single words, they can be of clusters 278.

See also: WordList display 318

Online step-by-step guide showing how

9.2 index

Explanations

Limitations 437
Summary Statistics 66
Match List 92
Mutual Information 289
Sort Order 315
Stop Lists 120
Type/token Ratios 303

Procedures

Auto-Join 273
Batch Processing 39
Calling up a Concordance 234
Choosing Texts 44
Colours 60
Computing a new variable 63

Folders 431
Editing Entries 72
Editing Filenames 113
Keyboard Shortcuts 439
Exiting 101
Fonts 78
Minimum & Maximum Settings 314
Mutual Information Score Computing 294
Printing 80
Re-sorting a Word List 315
Saving Results 101
Searching for an Entry by Typing 109
Searching for Entry-types using Menu 288
Single Words or Clusters 278
Text Characteristics 124
Word Index 276
Zapping entries 129

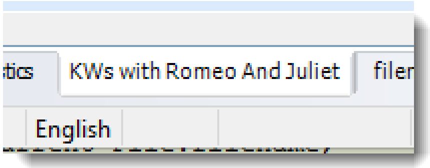
See also: WordSmith Main Index 2, WordList display 318

9.3 compare word lists

9.3.1 compute key words

With a word list visible in the WordList tool, you may choose *Compute* | *KeyWords* to get a keywords analysis of the current word list. This will assume you will wish to use the reference corpus 447 defined in the settings 254 for comparison.

You will see the results in one of the tabs at the bottom of the screen.



As in the **KeyWords** tool, this procedure compares all the words in your original word list with those in the reference corpus but does not inform you about words which are only found in the reference corpus.

See also : Compare two wordlists 260, word-list with tags as prefix 315

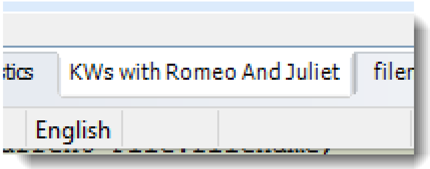
9.3.2 comparing wordlists

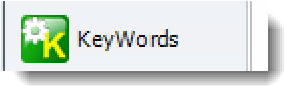
The idea is to help stylistic comparisons. Suppose you're studying several versions of a story, or different translations of it. If one version uses *kill* and another has *assassinate*, you can use this function.

The procedure compares *all the words in both lists* and will report on all those which appear significantly more often in one than the other, including those which appear more than a minimum number of times in one even if they do not appear at all in the other.

How to do it

1. Open a word list.
2. In the menu, choose *File* | *Compare 2 wordlists*.
3. Choose a word list to compare with. You will see the results in one of the tabs at the bottom of the screen.





The minimum frequency (which you can alter in the Controller 4, *KeyWords Settings* tab) can be set to 1. If it is raised to say 3, the comparison will ignore words which do not appear at least 3 times in at least one of the two lists.

Choose the significance value (all, or a p value 235 from 0.1 to 0.000001 or what you will). The smaller the p value 235, the more selective the comparison. In other words, a p setting of 0.1 will show more words than a p setting of 0.0001 will.

The display 261 format is similar to that used in KeyWords 229. You will also find the Dice coefficient 433 which compares the vocabularies of the two texts, reported in the Notes 29.

See also: Compute Key Words 259, Consistency Analysis 263, Match List 92

9.3.3 comparison display

How to get here? by choosing compare two wordlists 260

Here is a comparison window, where we have compared Shakespeare's King Lear with Romeo and Juliet.

The display shows

frequency in the text you started with, here *King Lear,* (with % if > 0.01%) -- then, to the right

frequency in the other text, here *Romeo & Juliet,* (with % if > 0.01%) -- then, to the right

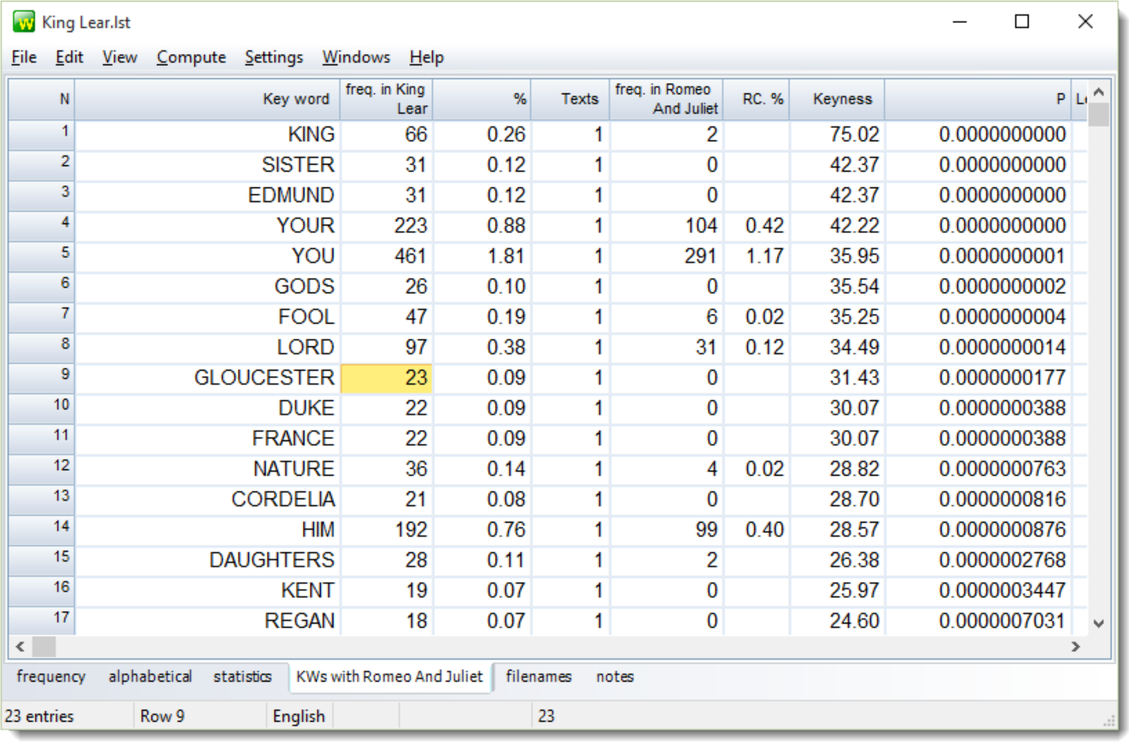
chi-square or log likelihood 245, and p value 235.

The criterion for what counts as "outstanding" is based on the minimum probability value entered before the lists were compared. The smaller this probability value the fewer words in the display.

The words appear sorted according to how outstanding their frequencies of occurrence are. Those near the top are outstandingly frequent in your main word-list. At the end of the listing you'll find those which are outstandingly infrequent in the first text chosen: in other words, key in the second text.

This comparison is similar to the analysis of "key words" in the KeyWords 229 program. The KeyWords analysis is slightly quicker and allows for batch processing.

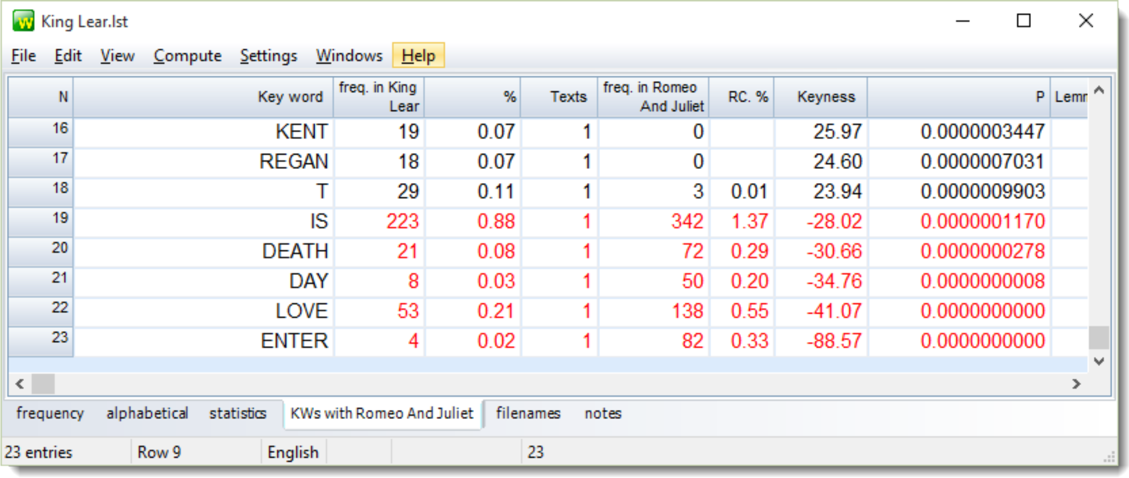
The word **`King`** is the most key of all, it scores 75 in the keyness column.



N	Key word	freq. in King Lear	%	Texts	freq. in Romeo And Juliet	RC. %	Keyness	P	L
1	KING	66	0.26	1	2		75.02	0.0000000000	
2	SISTER	31	0.12	1	0		42.37	0.0000000000	
3	EDMUND	31	0.12	1	0		42.37	0.0000000000	
4	YOUR	223	0.88	1	104	0.42	42.22	0.0000000000	
5	YOU	461	1.81	1	291	1.17	35.95	0.0000000001	
6	GODS	26	0.10	1	0		35.54	0.0000000002	
7	FOOL	47	0.19	1	6	0.02	35.25	0.0000000004	
8	LORD	97	0.38	1	31	0.12	34.49	0.0000000014	
9	GLOUCESTER	23	0.09	1	0		31.43	0.0000000177	
10	DUKE	22	0.09	1	0		30.07	0.0000000388	
11	FRANCE	22	0.09	1	0		30.07	0.0000000388	
12	NATURE	36	0.14	1	4	0.02	28.82	0.0000000763	
13	CORDELIA	21	0.08	1	0		28.70	0.0000000816	
14	HIM	192	0.76	1	99	0.40	28.57	0.0000000876	
15	DAUGHTERS	28	0.11	1	2		26.38	0.0000002768	
16	KENT	19	0.07	1	0		25.97	0.0000003447	
17	REGAN	18	0.07	1	0		24.60	0.0000007031	

23 entries | Row 9 | English | 23

At the bottom we see the words of *King Lear* which are least key in comparison with the play *Romeo and Juliet*.


King Lear.lst
File Edit View Compute Settings Windows Help
frequency alphabetical statistics KWs with Romeo And Juliet filenames notes
23 entries Row 9 English 23

N	Key word	freq. in King Lear	%	Texts	freq. in Romeo And Juliet	RC. %	Keyness	P	Lemr
16	KENT	19	0.07	1	0		25.97	0.0000003447	
17	REGAN	18	0.07	1	0		24.60	0.0000007031	
18	T	29	0.11	1	3	0.01	23.94	0.0000009903	
19	IS	223	0.88	1	342	1.37	-28.02	0.0000001170	
20	DEATH	21	0.08	1	72	0.29	-30.66	0.0000000278	
21	DAY	8	0.03	1	50	0.20	-34.76	0.0000000008	
22	LOVE	53	0.21	1	138	0.55	-41.07	0.0000000000	
23	ENTER	4	0.02	1	82	0.33	-88.57	0.0000000000	

9.4 merging wordlists

The point of it

You might want to merge 2 word lists (or concordances, mutual information lists etc.) with each other if making each one takes ages or if you are gradually building up a master word list or concordance based on a number of separate genres or text-types.

How to do it

With one word-list (or concordance) opened, choose *File* | *Merge with* and select another.

Be aware that...

Making a merged word list implies that each set of source texts was different. If you choose to merge 2 word lists both of which contained information about the same text file, WordSmith will do as you ask even though the information about the number of occurrences and of texts in which each word-type was found is (presumably) inaccurate.

Merging a list in English with another in Spanish: if you start with the one in Spanish, the one in English will be merged in and henceforth treated as if it were Spanish, eg. in sort order. Presumably if you try to merge one in English with one in Arabic (I've never tried) you should see all the forms but you would get different results merging the Arabic one into the English one (all the Arabic words would be treated as if they were English).

9.5 consistency

9.5.1 consistency analysis (range)

This function (termed "range" by Paul Nation) comes automatically with any word-list.

In any word-list you will see a column headed "Texts". This shows the number of texts each word occurred in (the maximum here being the total number of text-files used for the word-list).

The point of it...

The idea is to find out which words recur consistently in lots of texts of a given genre. For example, the word `consolidate` was found to occur in many of a set of business Annual Reports. It did not occur very often in each of them, but did occur much more consistently in the business reports than in a mixed set of texts.
Naturally, words like `the` are consistent across nearly all texts in English. (While working on a set of word lists to compare with business reports, I found one text without **`the`**. I also discovered that one of my texts was in Italian: but this wasn't the one without **`the`**! The culprit was an election results list, which contained lots of instances of **`Cons., Lab.`** and place names, but no instances of **`the`**.)
To analyse common grammar words like `the`, a consistency list may be very useful. Even so, you're likely to find some common lexical items recur surprisingly consistently.

To eliminate the commonly consistent words and find only those which seem to characterise your genre or sub-genre, you need to find out which are significantly consistent. Save your word list, then use it for comparison 260 with others in WordList, or using KeyWords. This way you can determine which are the significantly consistent words in your genre or sub-genre.

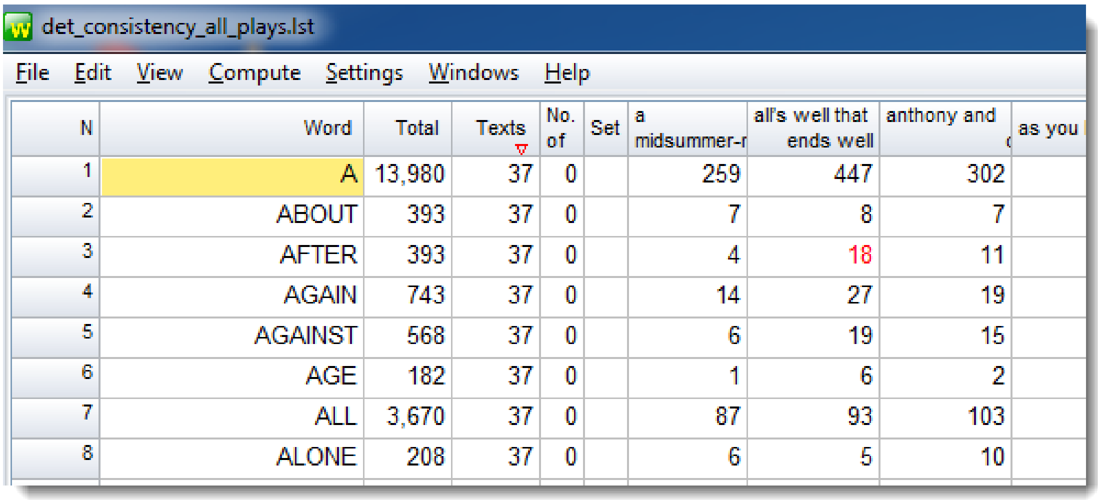
See also: Consistency Analysis (Detailed) 263, Comparing Word-lists 260, Match List 92

9.5.2 detailed consistency analysis

This function does exactly the same thing as simple consistency 262, but provides much more detail.

The point of it...

The idea is to help stylistic comparisons. Suppose you're studying several versions of a story, or different translations of it. This function enables you to see all the words which are used in the word lists which you have called up.


det_consistency_all_plays.lst
File Edit View Compute Settings Windows Help

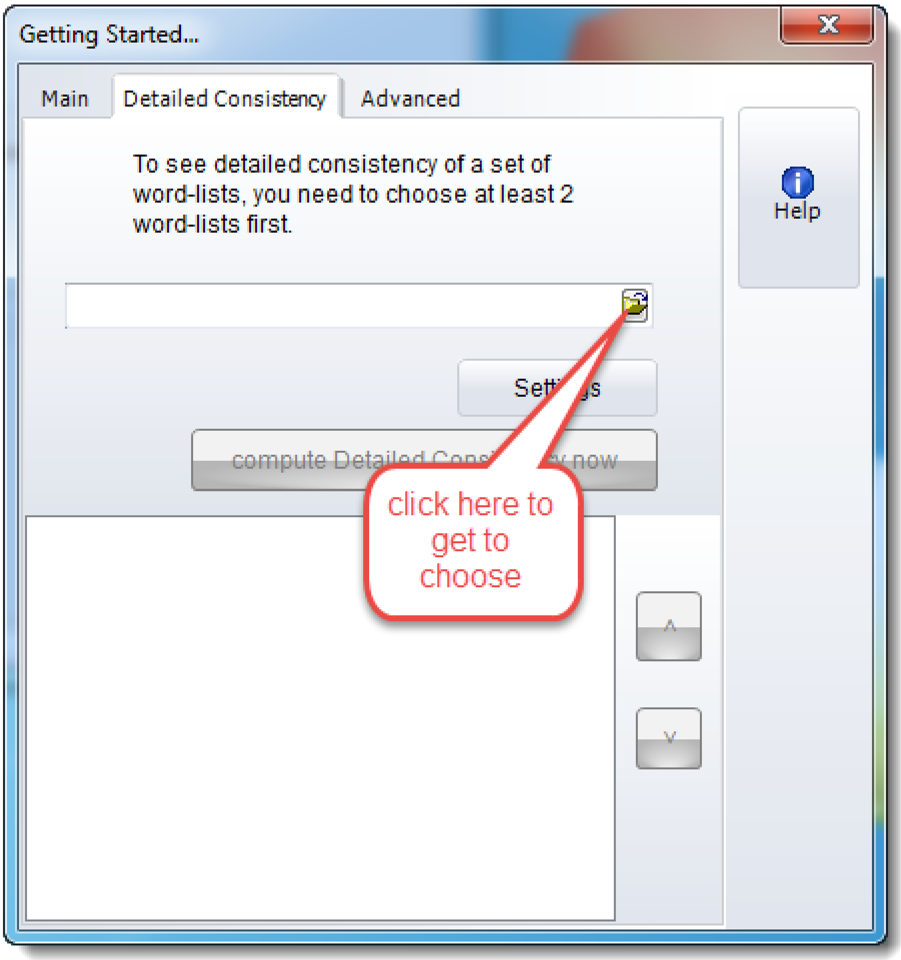
N	Word	Total	Texts	No. of	Set	a midsummer-r	all's well that ends well	anthony and	as you
1	A	13,980	37	0		259	447	302	
2	ABOUT	393	37	0		7	8	7	
3	AFTER	393	37	0		4	18	11	
4	AGAIN	743	37	0		14	27	19	
5	AGAINST	568	37	0		6	19	15	
6	AGE	182	37	0		1	6	2	
7	ALL	3,670	37	0		87	93	103	
8	ALONE	208	37	0		6	5	10	

The Total column shows how many instances of each word occurred overall, Texts shows how many text-files it came in. Then there are two columns (No. of Lemmas, and Set which behaves as in a word-list) and then a column for each text. In this case, the word **after** occurred in all 37 texts, it occurred 393 times in all, and it was most frequent in **all's well that ends well** at 18 occurrences. Statistics and filenames can be seen for the set of 37 Shakespeare plays used here by clicking on the tabs at the bottom. Notes 29 can be edited and saved along with the detailed consistency list.

There is no limit except the limit of available memory as to how many text files you can process in this procedure. You can set a minimum number of texts and a minimum overall frequency in the WordList settings in the Controller 311.

How to do it...

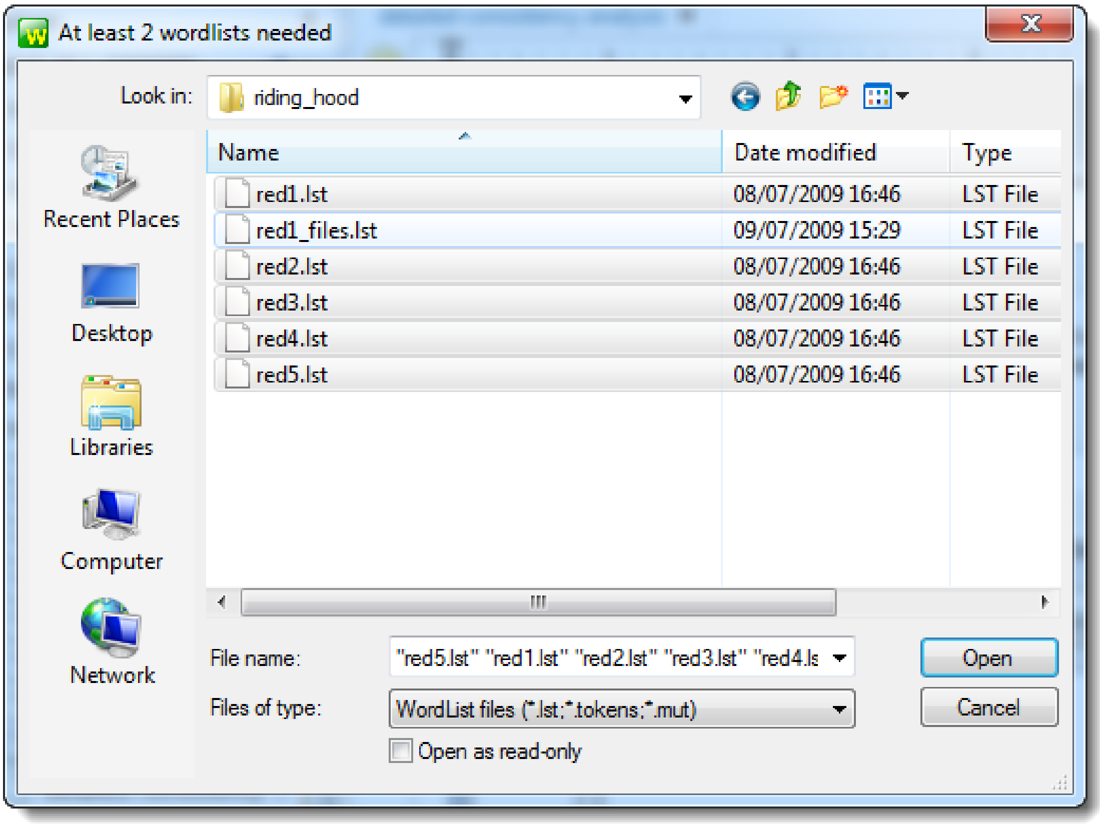
In the window you see when you press **New...(●)** you will be offered a tab showing detailed consistency.



Getting Started...

To see detailed consistency of a set of word-lists, you need to choose at least 2 word-lists first.

To choose more than 1, use Control or Shift as you click. Below I have chosen five out of 6 available. (These are versions of Red Riding Hood.)



Look in: riding_hood

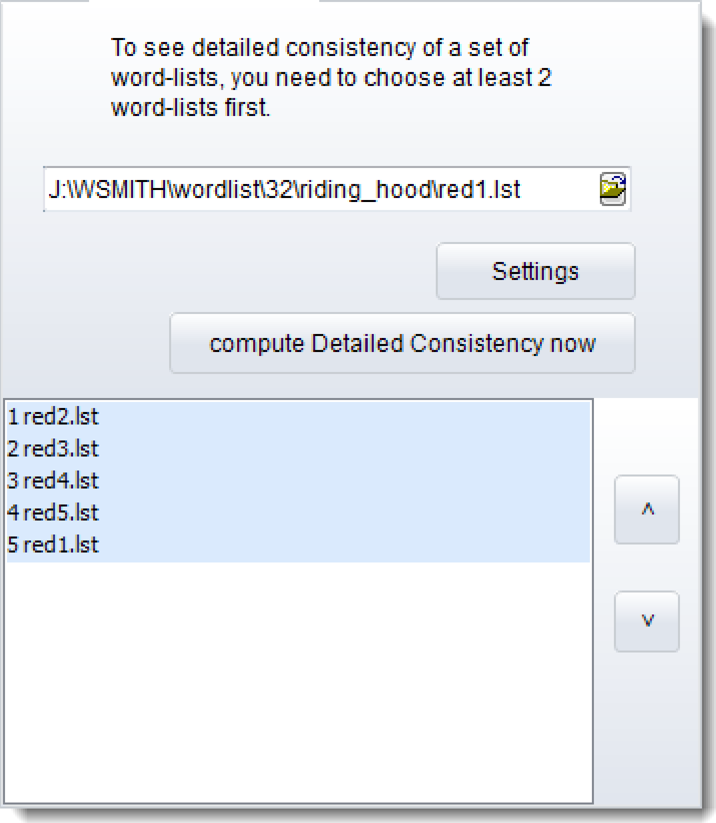
Name	Date modified	Type
red1.lst	08/07/2009 16:46	LST File
red1_files.lst	09/07/2009 15:29	LST File
red2.lst	08/07/2009 16:46	LST File
red3.lst	08/07/2009 16:46	LST File
red4.lst	08/07/2009 16:46	LST File
red5.lst	08/07/2009 16:46	LST File

File name: "red5.lst" "red1.lst" "red2.lst" "red3.lst" "red4.ls

Files of type: WordList files (*.lst;*.tokens;*.mut)

☐ Open as read-only

Initially they may come in the wrong order:

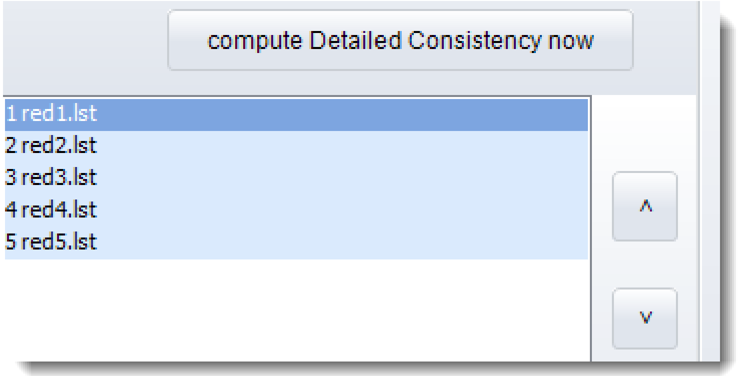


To see detailed consistency of a set of word-lists, you need to choose at least 2 word-lists first.

J:\WSMITH\wordlist\32\riding_hood\red1.lst

1 red2.lst
2 red3.lst
3 red4.lst
4 red5.lst
5 red1.lst

so adjust with the two buttons at the right.



1 red1.lst
2 red2.lst
3 red3.lst
4 red4.lst
5 red5.lst

and now press **compute Detailed Consistency now**.

Settings

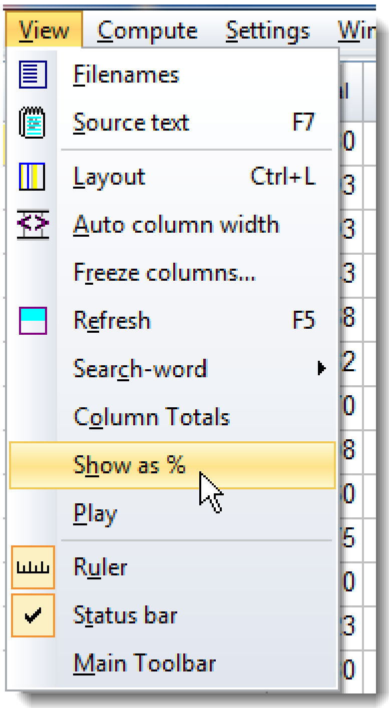
You can require a minimum number of texts and minimum frequency in the main Controller 311 if you click this.

Sorting

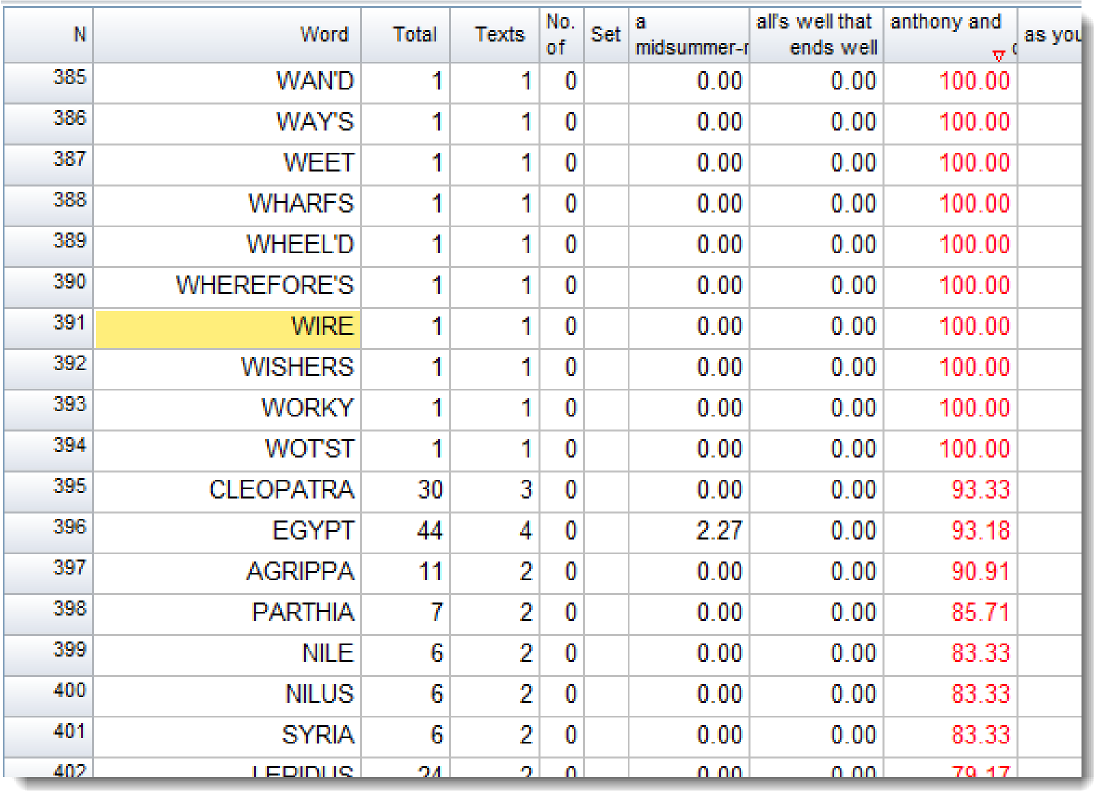
Each column can be sorted by clicking on its header column (**Word, Freq.** etc.). When working on Shakespeare plays, to get the words which occurred in all 37 to the top, I clicked **Texts**.

Row percentages

If you choose to *Show as %*,



you will transform the view so as to get row percentages. In this screenshot,



N	Word	Total	Texts	No. of	Set	a midsummer-r	all's well that ends well	anthony and	as you
385	WAN'D	1	1	0		0.00	0.00	100.00	
386	WAY'S	1	1	0		0.00	0.00	100.00	
387	WEET	1	1	0		0.00	0.00	100.00	
388	WHARFS	1	1	0		0.00	0.00	100.00	
389	WHEEL'D	1	1	0		0.00	0.00	100.00	
390	WHEREFORE'S	1	1	0		0.00	0.00	100.00	
391	WIRE	1	1	0		0.00	0.00	100.00	
392	WISHERS	1	1	0		0.00	0.00	100.00	
393	WORKY	1	1	0		0.00	0.00	100.00	
394	WOT'ST	1	1	0		0.00	0.00	100.00	
395	CLEOPATRA	30	3	0		0.00	0.00	93.33	
396	EGYPT	44	4	0		2.27	0.00	93.18	
397	AGRIPPA	11	2	0		0.00	0.00	90.91	
398	PARTHIA	7	2	0		0.00	0.00	85.71	
399	NILE	6	2	0		0.00	0.00	83.33	
400	NILUS	6	2	0		0.00	0.00	83.33	
401	SYRIA	6	2	0		0.00	0.00	83.33	
402	LEPIDUS	24	2	0		0.00	0.00	79.17	

we see the last few items which appear only in Anthony and Cleopatra, then Cleopatra (93.3%), Egypt (93.18%) etc. (Egypt appears also in A Midsummer Night's Dream, As You Like It, KIng Henry VIII.)

See also: Detailed Consistency Relations 268, Consistency Analysis (range) 262, Comparison Display 261, Comparing Word-lists 260, Match List 92, Column Totals 62

9.5.2.1 re-sorting: consistency lists

The frequency-ordered consistency display can be re-sorted by

alphabetical order (Word)

total frequencies overall (Total, the default)

by the *frequencies* in any given file (you see the file names).

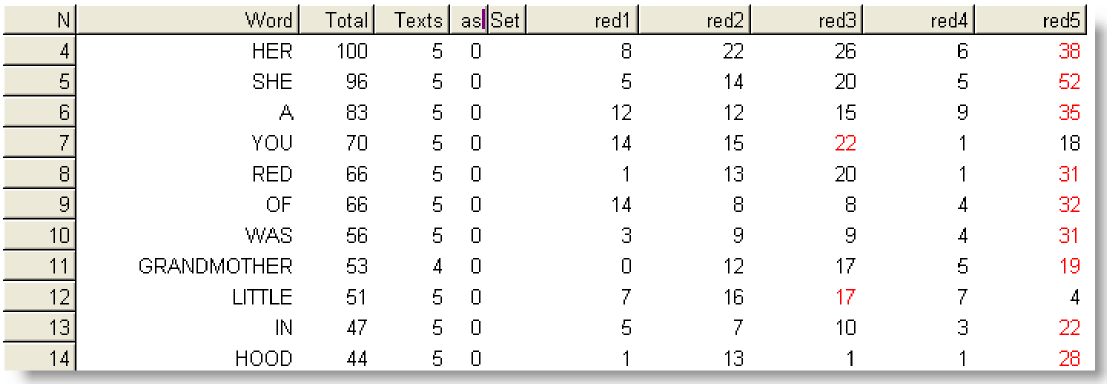
Click on Word, Total or a filename to choose.

The sort can be either ascending or descending, the default being descending.

See also: Sorting word-lists 315

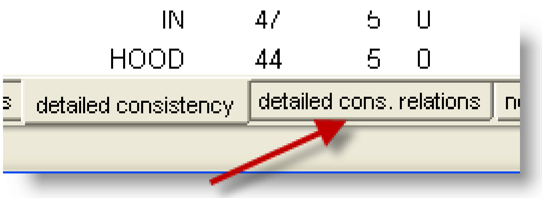
9.5.3 detailed consistency relations

With a detailed consistency list 263 such as this, of five versions of the fairy story *Little Red Riding Hood*,

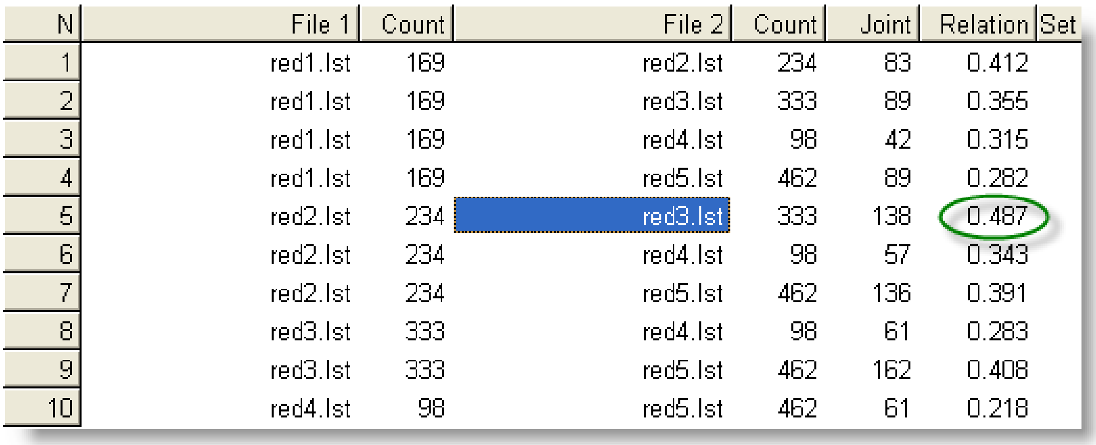


N	Word	Total	Texts	as	Set	red1	red2	red3	red4	red5
4	HER	100	5	0		8	22	26	6	38
5	SHE	96	5	0		5	14	20	5	52
6	A	83	5	0		12	12	15	9	35
7	YOU	70	5	0		14	15	22	1	18
8	RED	66	5	0		1	13	20	1	31
9	OF	66	5	0		14	8	8	4	32
10	WAS	56	5	0		3	9	9	4	31
11	GRANDMOTHER	53	4	0		0	12	17	5	19
12	LITTLE	51	5	0		7	16	17	7	4
13	IN	47	5	0		5	7	10	3	22
14	HOOD	44	5	0		1	13	1	1	28

it looks as if the most long-winded story is probably version 5 (**red5.lst**). If you click the *detailed cons. relation* tab



you can see the relevant statistics more usefully:



N	File 1	Count	File 2	Count	Joint	Relation	Set
1	red1.lst	169	red2.lst	234	83	0.412	
2	red1.lst	169	red3.lst	333	89	0.355	
3	red1.lst	169	red4.lst	98	42	0.315	
4	red1.lst	169	red5.lst	462	89	0.282	
5	red2.lst	234	red3.lst	333	138	0.487	
6	red2.lst	234	red4.lst	98	57	0.343	
7	red2.lst	234	red5.lst	462	136	0.391	
8	red3.lst	333	red4.lst	98	61	0.283	
9	red3.lst	333	red5.lst	462	162	0.408	
10	red4.lst	98	red5.lst	462	61	0.218	

where it can be seen that **red5** has a type-count of 462 words, more than any other, and that the relation between **red2** and **red3** is the closest with a relation statistic of 0.487.

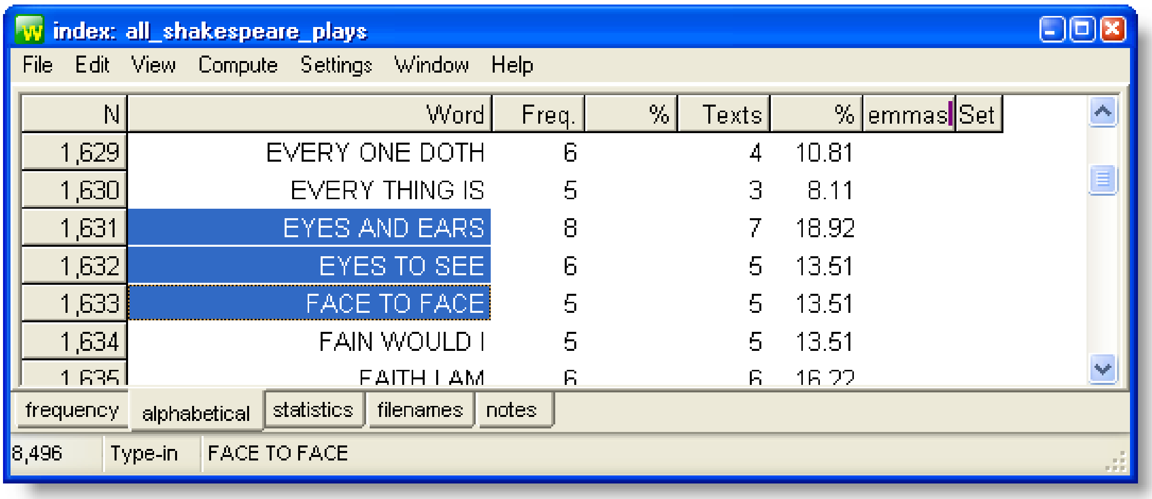
This relation is the Dice coefficient 433, based on the joint frequency and the type-counts of the two texts. Type count is the number of different word types 426 in each text. Joint frequency: there are 138 matches in the vocabulary of these two versions, which means that 138 distinct word types matched up in the two word lists. (If for example **book** appeared 20 times in one list and 3 times in the other, that would count as 1 match.)

A Dice coefficient ranges between 0 and 1. The 0.487 can be thought of like a percentage, i.e. there's about a 49% overlap between the vocabularies of the two versions of the same story.

See also : Detailed Consistency 263.

9.6 find filenames

If you have an index-based word list on screen you can see how many text files each word was found in. For example, in this index based on Shakespeare plays, **EYES AND EARS** occurs in 7 of the 37 plays.

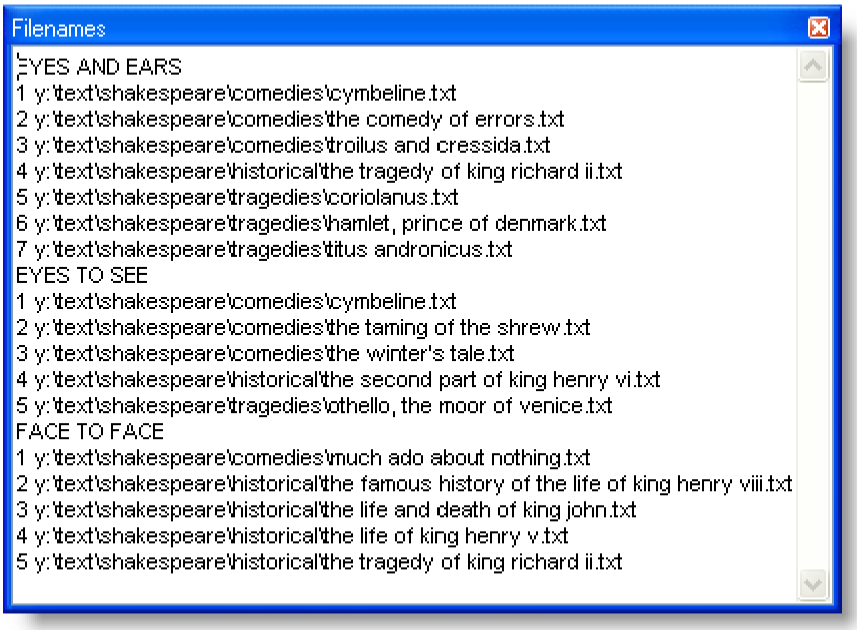


N	Word	Freq.	%	Texts	%	emmas	Set
1,629	EVERY ONE DOTH	6		4	10.81		
1,630	EVERY THING IS	5		3	8.11		
1,631	EYES AND EARS	8		7	18.92		
1,632	EYES TO SEE	6		5	13.51		
1,633	FACE TO FACE	5		5	13.51		
1,634	FAIN WOULD I	5		5	13.51		
1,635	FAITH I AM	6		6	16.22		

8,496 Type-in FACE TO FACE

What if you want to know *which* of those plays?

Select the word(s) or cluster(s) you're interested in and choose *File* | *Find Files* in the menu and you will get something like this:



```
EYES AND EARS
1 y:\text\shakespeare\comedies\cymbeline.txt
2 y:\text\shakespeare\comedies\the comedy of errors.txt
3 y:\text\shakespeare\comedies\troilus and cressida.txt
4 y:\text\shakespeare\historical\the tragedy of king richard ii.txt
5 y:\text\shakespeare\tragedies\coriolanus.txt
6 y:\text\shakespeare\tragedies\hamlet, prince of denmark.txt
7 y:\text\shakespeare\tragedies\titus andronicus.txt
EYES TO SEE
1 y:\text\shakespeare\comedies\cymbeline.txt
2 y:\text\shakespeare\comedies\the taming of the shrew.txt
3 y:\text\shakespeare\comedies\the winter's tale.txt
4 y:\text\shakespeare\historical\the second part of king henry vi.txt
5 y:\text\shakespeare\tragedies\othello, the moor of venice.txt
FACE TO FACE
1 y:\text\shakespeare\comedies\much ado about nothing.txt
2 y:\text\shakespeare\historical\the famous history of the life of king henry viii.txt
3 y:\text\shakespeare\historical\the life and death of king john.txt
4 y:\text\shakespeare\historical\the life of king henry v.txt
5 y:\text\shakespeare\historical\the tragedy of king richard ii.txt
```

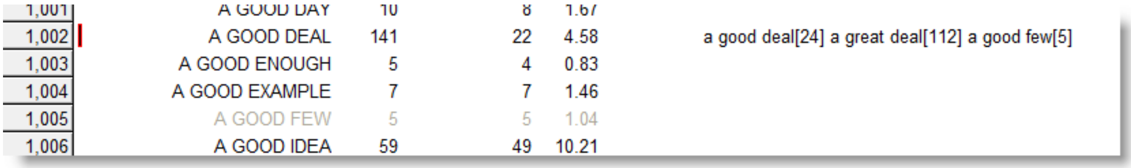
See also : source texts 116, selecting multiple entries 111, making a WordList index 276

9.7 Lemmas (joining words)

9.7.1 what are lemmas and how do we join words?

In a word list, a key word list or a list of collocates you may want to store several entries together: e.g. **`want; wants; wanting; wanted.`** Bringing them together means you're treating them as members of the same "lemma" or set -- rather like a headword in a dictionary.

A lemmatised head entry has a red mark in the left margin beside it. The others you marked will be coloured as if deleted. The linked entries which have been joined to the head can be seen at the right.



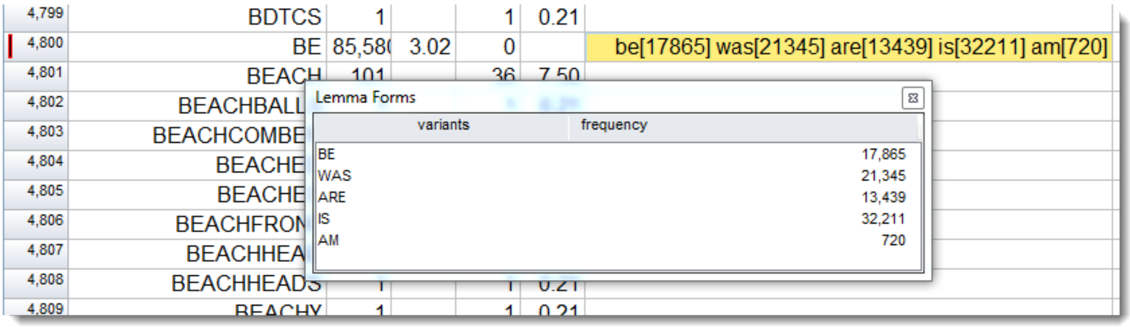
1,001	A GOOD DAY	10	8	1.67	
1,002	A GOOD DEAL	141	22	4.58	a good deal[24] a great deal[112] a good few[5]
1,003	A GOOD ENOUGH	5	4	0.83	
1,004	A GOOD EXAMPLE	7	7	1.46	
1,005	A GOOD FEW	5	5	1.04	
1,006	A GOOD IDEA	59	49	10.21	

Here we see a word list based on 3-word clusters 278 where originally **`a good deal`** had a frequency of 24, but has been joined to **`a great deal`** and **`a good few`** and thereby risen to 141.

Joining can be done automatically 273 or manually 271.

View all the various lemma forms

Double-click on the Lemmas column as in the shot below,



4,799	BDTCS	1		1	0.21	
4,800	BE	85,58(	3.02	0		be[17865] was[21345] are[13439] is[32211] am[720]
4,801	BEACH	101		36	7.50	
4,802	BEACHBALL					
4,803	BEACHCOMBE					
4,804	BEACHE					
4,805	BEACHE					
4,806	BEACHFRON					
4,807	BEACHHEA					
4,808	BEACHHEADS	1		1	0.21	
4,809	BEACHY	1		1	0.21	

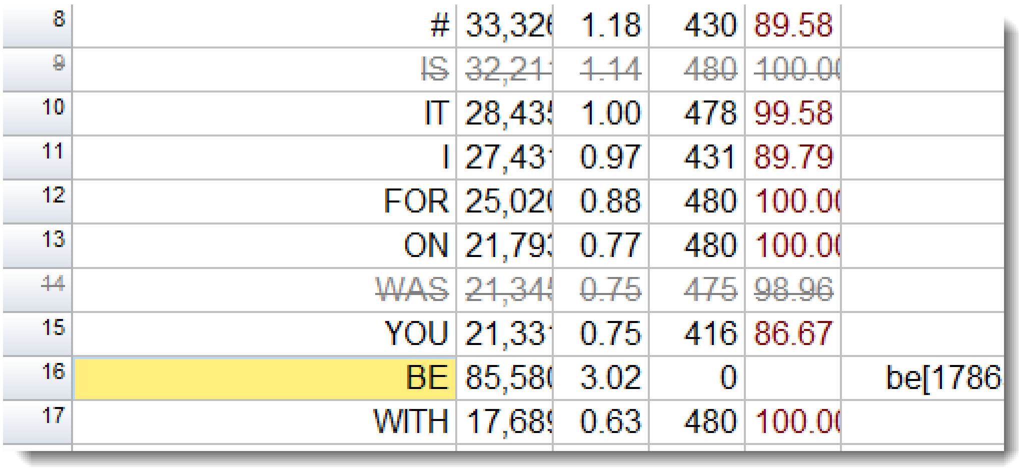
Lemma Forms

variants	frequency
BE	17,865
WAS	21,345
ARE	13,439
IS	32,211
AM	720

and a window of Lemma Forms will open up, showing the various components.

Get rid of the deleted words

If you don't want to see the deleted words



8	#	33,32(	1.18	430	89.58	
~~9~~	~~IS~~	~~32,21~~	~~1.14~~	~~480~~	~~100.0(~~	
10	IT	28,43!	1.00	478	99.58	
11	I	27,43	0.97	431	89.79	
12	FOR	25,02(	0.88	480	100.0(	
13	ON	21,79	0.77	480	100.0(	
~~14~~	~~WAS~~	~~21,34~~	~~0.75~~	~~475~~	~~98.96~~	
15	YOU	21,33	0.75	416	86.67	
16	BE	85,58(	3.02	0		be[1786
17	WITH	17,68!	0.63	480	100.0(	

choose Ctrl-Z to zap 129 them.

See also: Auto-Joining methods 273, Using a text file to lemmatise 274, selecting multiple entries 111, Concord lemmatisation 187

9.7.2 manual joining

Manual joining

You can simply do this by dragging one entry to another. Suppose your word list has

```
WANT
WANTED
WANTING
```

you can simply grab **wanting** or **wanted** with your mouse and place it on **want**.

(See choosing lemma file 274 if you want to join these to a word which isn't in the list)

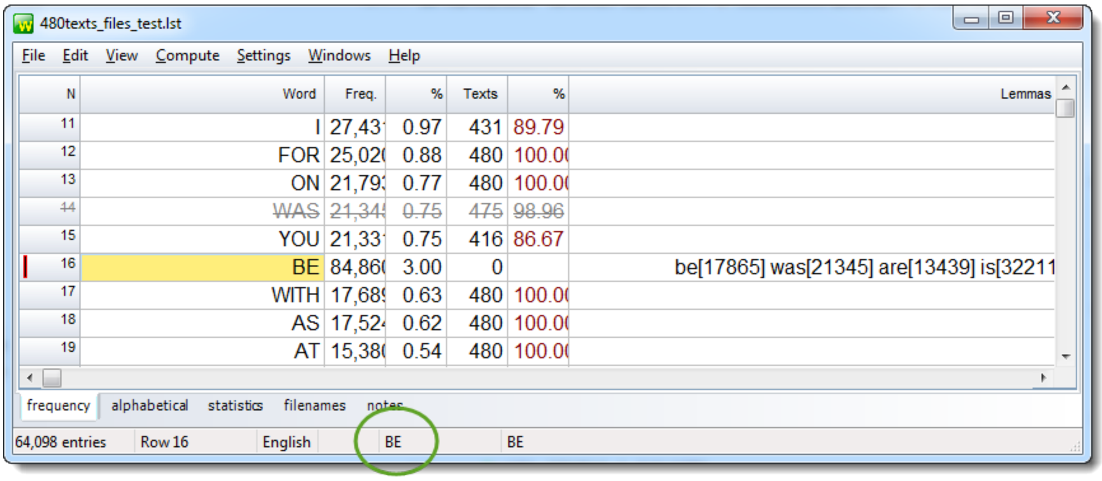
Can't see the word to join to?

If you cannot see all the items you want to join in one screen, you can do the same thing using by marking 112.

1. Use Alt+F5 to mark an entry for joining to another. The first one you mark will be the "head". For the moment, while you're still deciding which other entries belong with it, the edge of that row will be marked green. Any entries which you then decide to link with the head (by again pressing Alt+F5) will show they're marked too, in white. (If you change your mind you can press Shift+Alt+F5 and the marking will disappear.)

2. Use F4 to join all the entries which you've marked. The program will then put the joint frequencies of all the words you've marked with the frequency of the one you marked 112 first (the head).

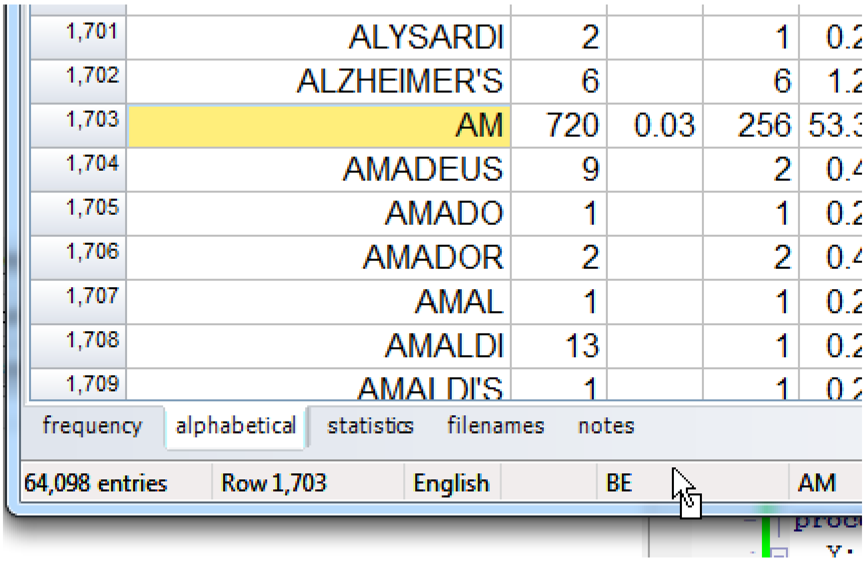
Alternatively, 1. select the head word, this makes it visible in the status bar.


480texts_files_test.lst
File Edit View Compute Settings Windows Help
frequency alphabetical statistics filenames notes

N	Word	Freq.	%	Texts	%	Lemmas
11	I	27,43	0.97	431	89.79	
12	FOR	25,02	0.88	480	100.0	
13	ON	21,79	0.77	480	100.0	
~~14~~	~~WAS~~	~~21,34~~	~~0.75~~	~~475~~	~~98.96~~	
15	YOU	21,33	0.75	416	86.67	
16	BE	84,86	3.00	0		be[17865] was[21345] are[13439] is[32211
17	WITH	17,68	0.63	480	100.0	
18	AS	17,52	0.62	480	100.0	
19	AT	15,38	0.54	480	100.0	

64,098 entries | Row 16 | English | BE | BE

2. Find the word you want to join and drag it to the status bar.



1,701	ALYSARDI	2		1	0.2
1,702	ALZHEIMER'S	6		6	1.2
1,703	AM	720	0.03	256	53.3
1,704	AMADEUS	9		2	0.4
1,705	AMADO	1		1	0.2
1,706	AMADOR	2		2	0.4
1,707	AMAL	1		1	0.2
1,708	AMALDI	13		1	0.2
1,709	AMALDI'S	1		1	0.2

64,098 entries | Row 1,703 | English | BE | AM

To Un-join

If you select an item which has lemmas visible at the right and press Ctrl+F4, this will unjoin the entries of that one lemma. To unjoin all lemmatised forms in the entire list, in the menu choose *Edit | Join | Unjoin All*.

9.7.3 auto-joining lemmas

There are two methods, a) based on a list, and b) based on a template.

a) File-based joining

You can join up lemmas using a text file 274 which automates the matching & joining process. The actual processing of the list takes place when you choose the menu option *Match Lemmas* (☰) in WordList, Concord or KeyWords. Every entry in your lemma list will be checked to see whether it matches one of the entries in your word list. In the example, if, say, ***am, was***, and ***were*** are found, they will be stored as lemmas of ***be***. If ***go*** and ***went*** are found, then ***went*** will be joined to ***go***.

b) Auto-joining based on a template

Or you can auto-join any of the entries in your current word list which meet your criteria: the menu option *Auto-Join* can be used to specify a string such as **S** or **S;ED;ING** and will then go through the whole word list, lemmatising all entries where one word only differs from the next by having **S** or **ED** or **ING** on the end of it. (Use **;** to separate multiple suffixes.)

Prefix / Suffix / Infix

By default all strings typed in are assumed to be suffixes; to join prefixes put an asterisk (*****) at the right end of the prefix. If you want to search for infixes (eg. **bloody** in **absobloodylutely** [languages like Swahili use infixes a lot]) put an asterisk at each end.

Examples

S;ED;ING will join **books** to **book, booked** to **book** and **booking** to **book**
***S;*ED;*ING** will join **books** to **book, booked** to **book** and **booking** to **book**
UN*;ED;ING will join **undo** to **do, booked** to **book** and **booking** to **book**
BLOODY will join **absobloodylutely** to **absolutely**

The process can be left to run quickly and automatically, or you can have it confirm with you before joining each one. Automatic lemmatisation, like search-and-replace spell-checking, can produce oddities if just left to run!
To stop in the middle of auto-joining, press Escape.

Tip

With a previously saved list, try auto-joining *without* confirming the changes (or choose *Yes to All* during it). Then choose the Alphabetical (as opposed to Frequency) version of the list and sort on Lemmas (by pressing the *Lemmas* column heading). You will see all the joined entries at the top of the list. It may be easier to Unjoin 270 (Ctrl+F4) any mistakes than to confirm each one... Finally, sort on the *Word* and save.

See also: Lemmatisation 270

9.7.4 choosing lemma file

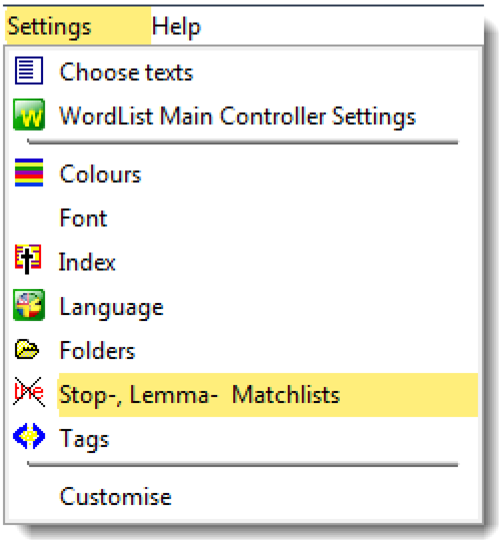
The point of it...

You may choose to lemmatise all items in the current word-list using a standard text file which groups words which belong together (**`be -> was, is, were`**, etc.). While it is time-consuming producing the text file the first time, it will be very useful if you want to lemmatise lots of word lists, and is much less "hit-and-miss" than auto-joining 273 using a template.

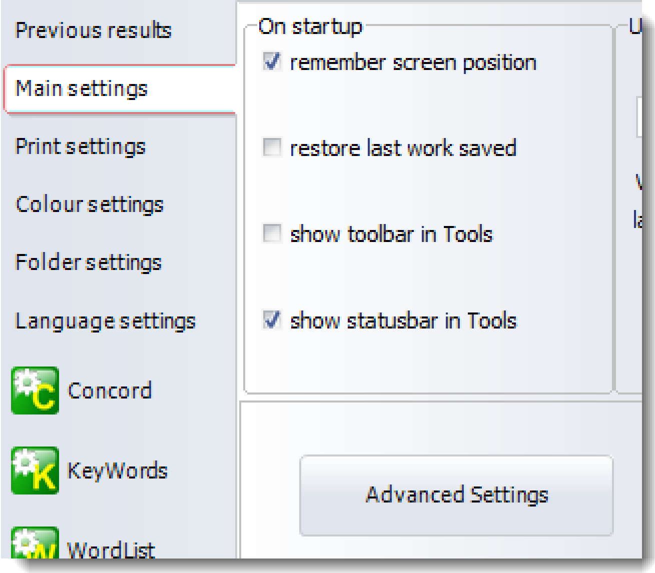
There is an English-language lemma list from Yasumasa Someya at http://lexically.net/downloads/BNC_wordlists/e_lemma.txt.

How to do it

Lemma list settings are accessed via the Lists option in the WordList menu

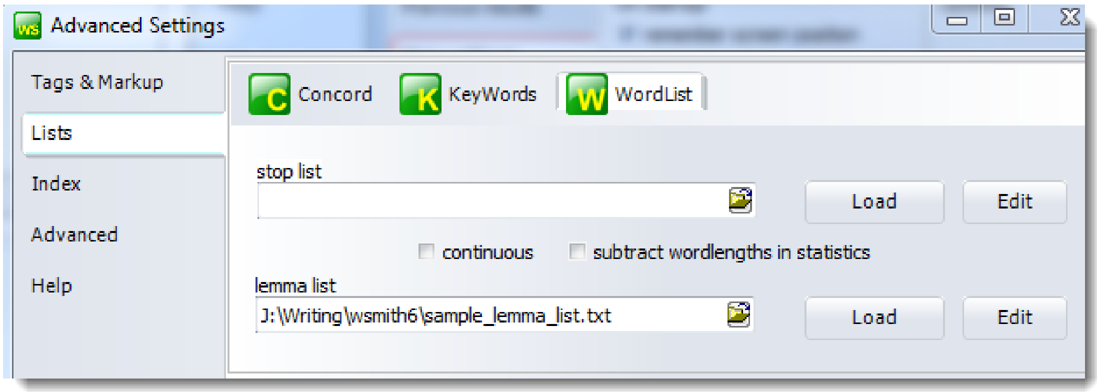


or an Advanced Settings button in the Controller



On startup
☑ remember screen position
□ restore last work saved
□ show toolbar in Tools
☑ show statusbar in Tools

followed by



stop list
□ continuous □ subtract wordlengths in statistics
lemma list J:\Writing\wsmith6\sample_lemma_list.txt

Choose the appropriate button (for Concord, KeyWords or WordList) and type the file name or browse for it, then Load it.

The file should contain a plain text list of lemmas with items like this:

```
BE -> AM, ARE, WAS, WERE, IS
GO -> GOES, GOING, GONE, WENT
```

WordSmith then reads the file and displays them (or a sample if the list is long). The format allows any alphabetic or numerical characters in the language the list is for, plus the single apostrophe, space, underscore. In other words, if you mistakenly put **GO = GOES** that line won't be included because of the = symbol.

The actual processing of the list will take place when you compute your word list, key word list or concordance or when you choose the menu option *Match Lemmas* (☰) in WordList, Concord or KeyWords. See Match List 92 for a more detailed explanation, with screenshots. Lemmatising occurs before any stop list 120 is processed.

What if my text files don't contain the headword of the lemma?

Suppose you are matching **AM, ARE** etc with **BE** as in the list above, but your texts don't actually contain the word **BE**. In that case the tool will insert **BE** with zero frequency and add **AM, ARE** etc as needed.

See also: Lemmatisation 270, Match List 92, Stop List 120, Lemmatisation in Concord 183

9.8 WordList Index

9.8.1 what is an Index for?

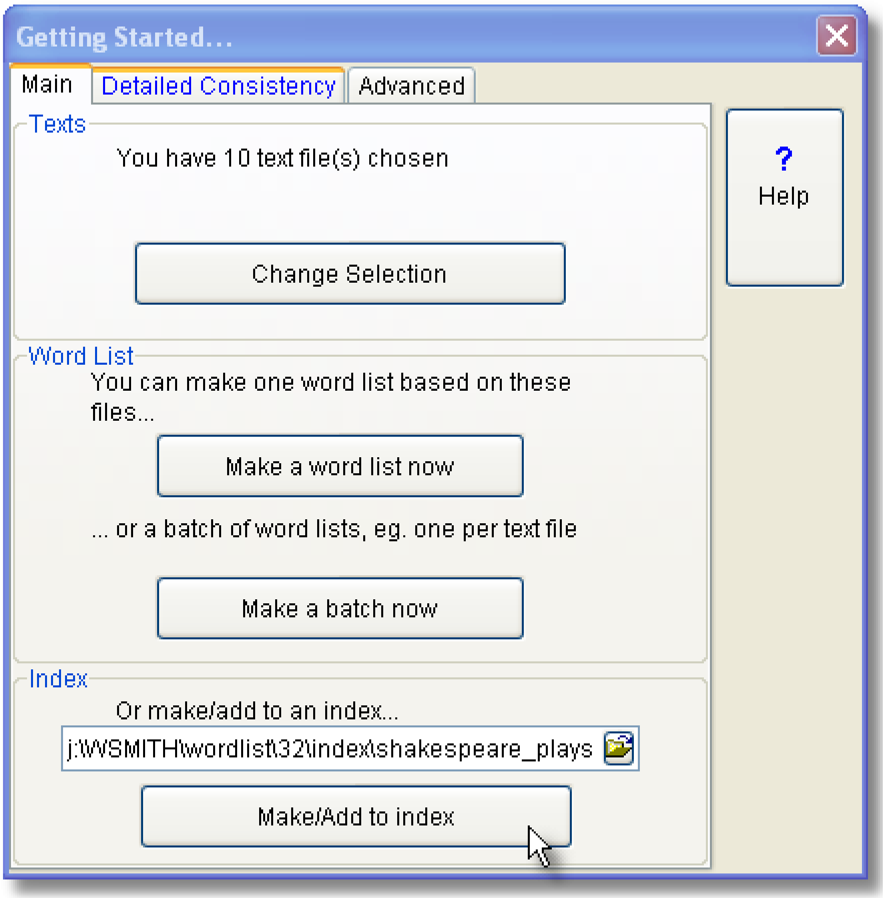
the point of it

1. One of the uses for an Index is to record the positions of all the words in your text file, so that you can subsequently see which word came in which part of each text. Another is to speed up access to these words, for example in concordancing. If you select one or more words in the index and press C, you get a speedy concordance.
2. Another is to compute "Mutual Information" 289 scores which relate word types to each other.
3. Or you can use an index to see word clusters 278.
4. Finally, an index is needed to generate concgram 12 searches.

See also Making an Index List 276, Viewing Index Lists 284, Exporting index data 286, find filenames for word clusters 269, WordList Help Contents 258, WSConcgram 12

9.8.2 making a WordList Index

The process is just like the one for making a word-list except that after choosing your texts and ensuring you like the index filename, you choose the bottom button here:



Getting Started...

Texts

You have 10 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

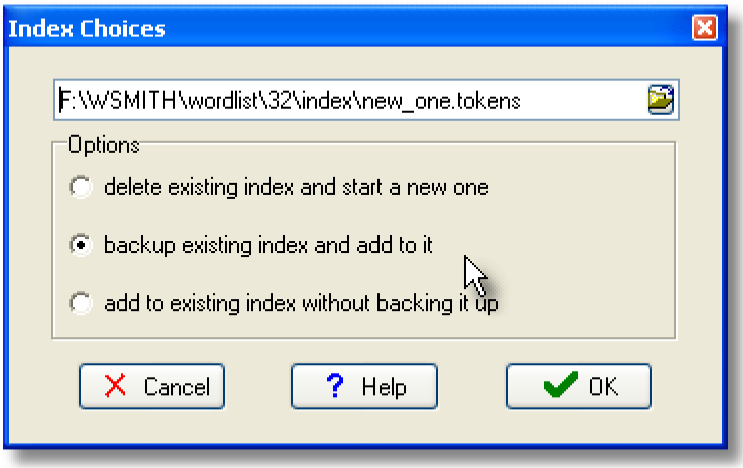
Index

Or make/add to an index...

j:\WSMITH\wordlist\32\index\shakespeare_plays

In this screenshot above, the basic filename is **`shakespeare_plays`**: WordSmith will add **`.tokens`** and **`.types`** to this basic filename as it works. Two files are created for each index:
`.tokens` file: a large file containing information about the position of every word token in your text files.
`.types` file: knows the individual word types.

If you choose an existing basic filename which you have already used, **WordList** will check whether you want to add to it or start it afresh:



Index Choices

F:\WSMITH\wordlist\32\index\new_one.tokens

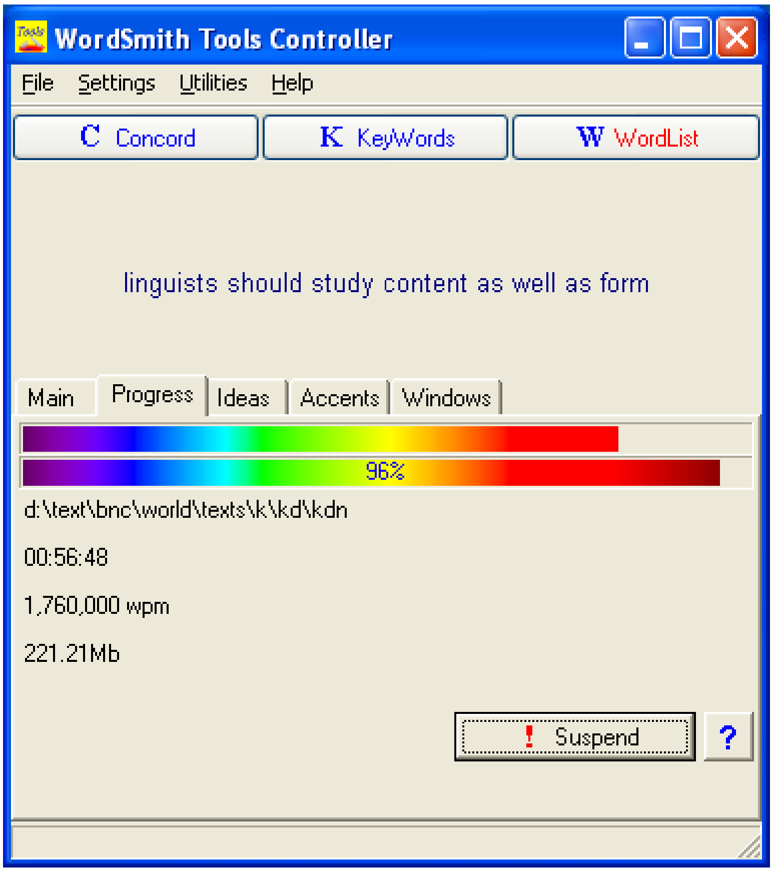
Options

○ delete existing index and start a new one

● backup existing index and add to it

○ add to existing index without backing it up

An index permits the computation of word clusters[278] and Mutual Information[289] scores for each word type. The screenshot below shows the progress bars for an index of the BNC corpus; on a modern PC it might work at a rate of about 2.8 million words per minute. The resulting **`BNC.tokens`** file was 1.6GB in size and the **`BNC.types`** file was 26 MB.



linguists should study content as well as form

d:\text\bnc\world\texts\k\kd\kdn

00:56:48

1,760,000 wpm

221.21Mb

adding to an index

To add to an existing index, just choose some more texts and choose *File* | *New* | *Index*. If the existing file-name is already in use for an index, you will be asked whether to add more or start it afresh as shown above.

See also Using Index Lists 276, Viewing Index Lists 284, WordList Help Contents 258.

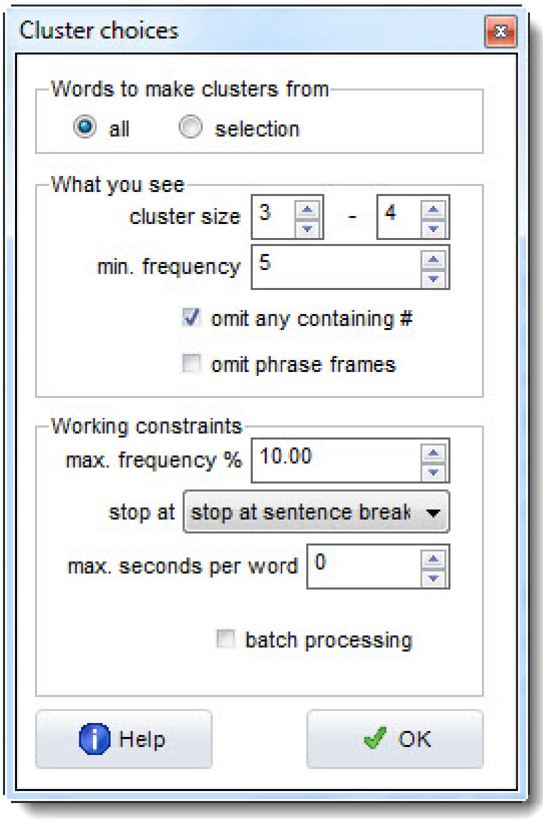
9.8.3 index clusters

WordList clusters

A word list doesn't need to be of single words. You can ask for a word list consisting of two, three, up to eight words on each line. To do cluster processing in WordList, first make an index 276.

How to see clusters...

Open 284 the index. Now choose *Compute* | *Clusters*.



Cluster choices

Words to make clusters from
● all ○ selection

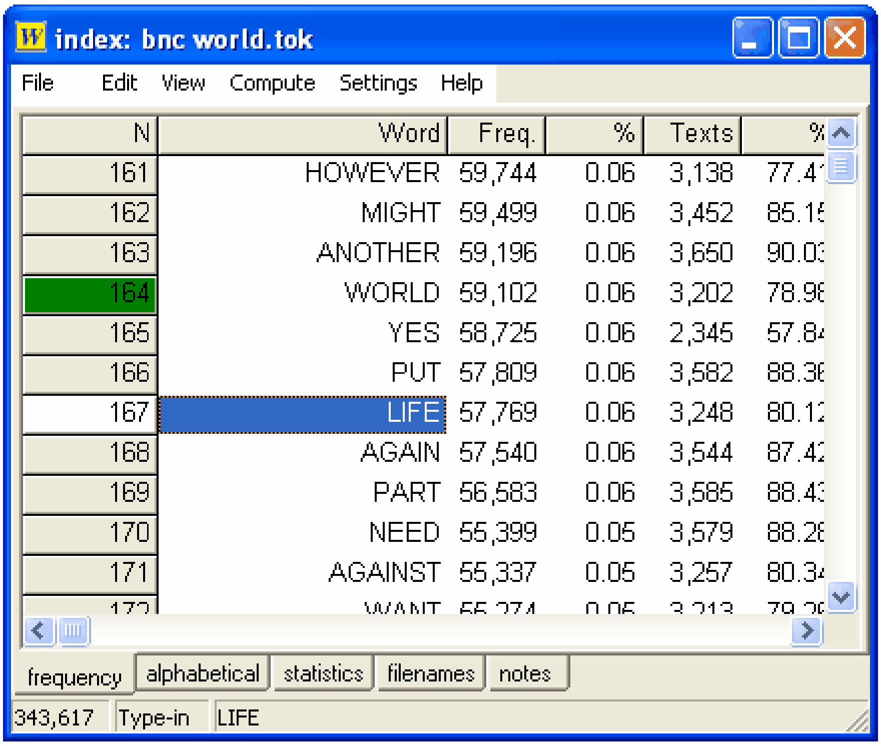
What you see
cluster size 3 - 4
min. frequency 5
☑ omit any containing #
□ omit phrase frames

Working constraints
max. frequency % 10.00
stop at: stop at sentence break
max. seconds per word 0
□ batch processing

Words to make clusters from

- "all" : all the clusters involving all words above a certain frequency (this will be s-l-o-w for a big corpus like the BNC), or
- "selection": clusters only for words you've selected (eg. you have highlighted BOOK and BOOKS and you want clusters like **`book a table, in my book`**).

  To choose words which aren't next to each other, press Control and click in the number at the left -- keep Control held down and click elsewhere. The first one clicked will go green and the others white. In the picture below, using an index of the BNC corpus, I selected **`world`** and then **`life`** by clicking numbers 164 and 167.



N	Word	Freq.	%	Texts	%
161	HOWEVER	59,744	0.06	3,138	77.41
162	MIGHT	59,499	0.06	3,452	85.15
163	ANOTHER	59,196	0.06	3,650	90.03
164	WORLD	59,102	0.06	3,202	78.98
165	YES	58,725	0.06	2,345	57.84
166	PUT	57,809	0.06	3,582	88.36
167	LIFE	57,769	0.06	3,248	80.12
168	AGAIN	57,540	0.06	3,544	87.42
169	PART	56,583	0.06	3,585	88.43
170	NEED	55,399	0.05	3,579	88.28
171	AGAINST	55,337	0.05	3,257	80.34
172	WANT	55,274	0.05	3,213	79.26

The process will take time. In the case of BNC, the index knows the positions of all of the 100 million words. To find 3-word clusters, in the case above, it took about a minute to process all the 115,000 cases of **world** and **life** and find 5,719 clusters like **the world bank** and **of real life**. Chris Tribble tells me it took his PC 36 hours to compute all 3-word clusters on the whole BNC ... he was able to use the PC in the meantime but that's not a job you're going to want to do often.

What you see

The *cluster size* must be between 2 and 8 words.
The *min. frequency* is the minimum number of each that you want to see.
omit #: if selected, this won't show any clusters involving numbers and dates
omit phrase frames: see phrase frames section below.
Here the user has chosen to see any 3-4-word clusters that appear 5 or more times.

Working constraints

The "max. frequency %" setting is to speed the process up.

in more detail...

It means the maximum frequency percentage which the calculation of clusters for a given word will process. This is because there are lots and lots of the very high frequency items and you may well not be interested in clusters which *begin* with them. For example, the item **the** is likely to be about 6% of any word-list (about 6 million of them in the BNC therefore), and you might not want clusters starting **the...** -- if so, you might set the max. percent to 0.5% or 0.1% (which for the BNC corpus will cut out the top 102 frequency words). You will still get clusters which include very high frequency items in the middle or end, like the **a** in **book a table**, but would not get **in my book**, which begins with the very high frequency word **in**. The more words you include, the longer the process will take....

Stop at, like Concord clusters[175], offers a number of constraints, such as sentence and other punctuation-marked breaks[188]. The idea is that a 5-word cluster which starts in one sentence and continues in the next is not likely to make much sense.

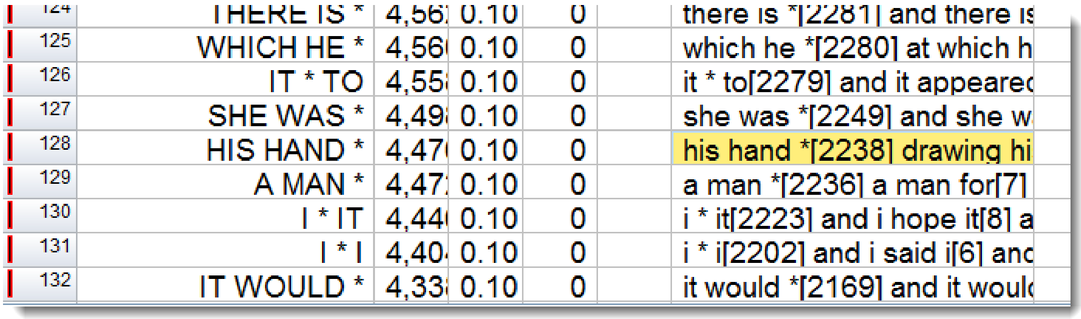
Max. seconds per word is another way of controlling how long the process will take. The default (0) means no limit. But if you set this e.g. to 30 then as WordList processes the words in order, as soon as one has taken 30 seconds no further clusters will be collected starting with that word.

batch processing allows you to create a whole set of cluster word-lists at one time.

Phrase frames

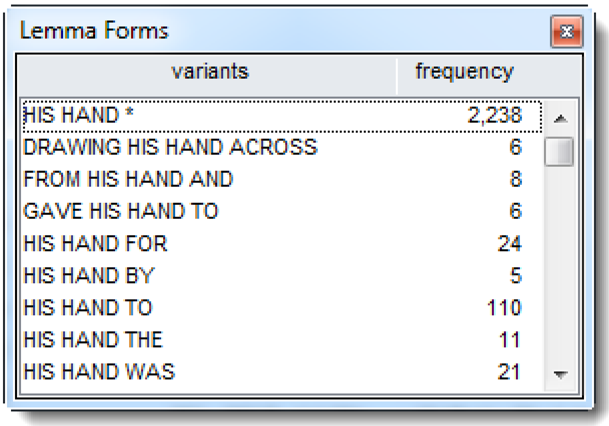
These are what William H. Fletcher has defined as *phrase-frames*, i.e. "groups of wordgrams identical but for a single word", in his kfNgram program.

Here, processing 23 Dickens novels shows lots of phrase frames where the wildcard word is represented with *.



124	THERE IS *	4,56:	0.10	0		there is *[2281] and there is
125	WHICH HE *	4,56(	0.10	0		which he *[2280] at which h
126	IT * TO	4,55:	0.10	0		it * to[2279] and it appearec
127	SHE WAS *	4,49:	0.10	0		she was *[2249] and she w
128	HIS HAND *	4,47(	0.10	0		his hand *[2238] drawing hi
129	A MAN *	4,47:	0.10	0		a man *[2236] a man for[7]
130	I * IT	4,44(	0.10	0		i * it[2223] and i hope it[8] a
131	I * I	4,40.	0.10	0		i * i[2202] and i said i[6] anc
132	IT WOULD *	4,33:	0.10	0		it would *[2169] and it woulc

If you double-click the lemmas column (highlighted here in yellow), you get to see the detail.



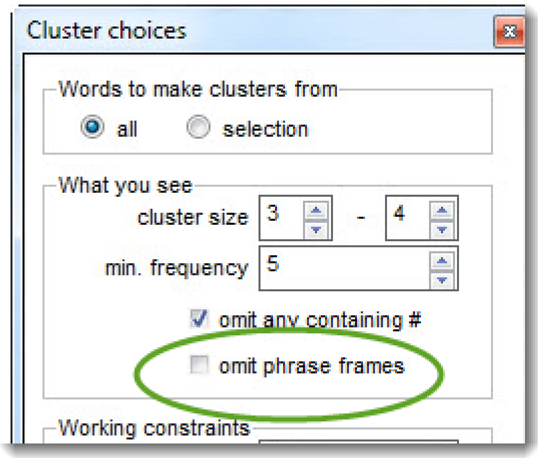
Lemma Forms

variants	frequency
HIS HAND *	2,238
DRAWING HIS HAND ACROSS	6
FROM HIS HAND AND	8
GAVE HIS HAND TO	6
HIS HAND FOR	24
HIS HAND BY	5
HIS HAND TO	110
HIS HAND THE	11
HIS HAND WAS	21

The process joins all the variants of the phrase in the Lemmas column. In the word list itself they will appear deleted (because they have been joined to another item, the phrase frame). You can un-join them all if you want (*Edit* | *Joining* | *Unjoin* or *Unjoin all*).

Omit phrase frames?

If you don't want to see phrase frames, select the *omit phrase frames* option.



Cluster choices

Words to make clusters from

● all ○ selection

What you see

cluster size 3 - 4

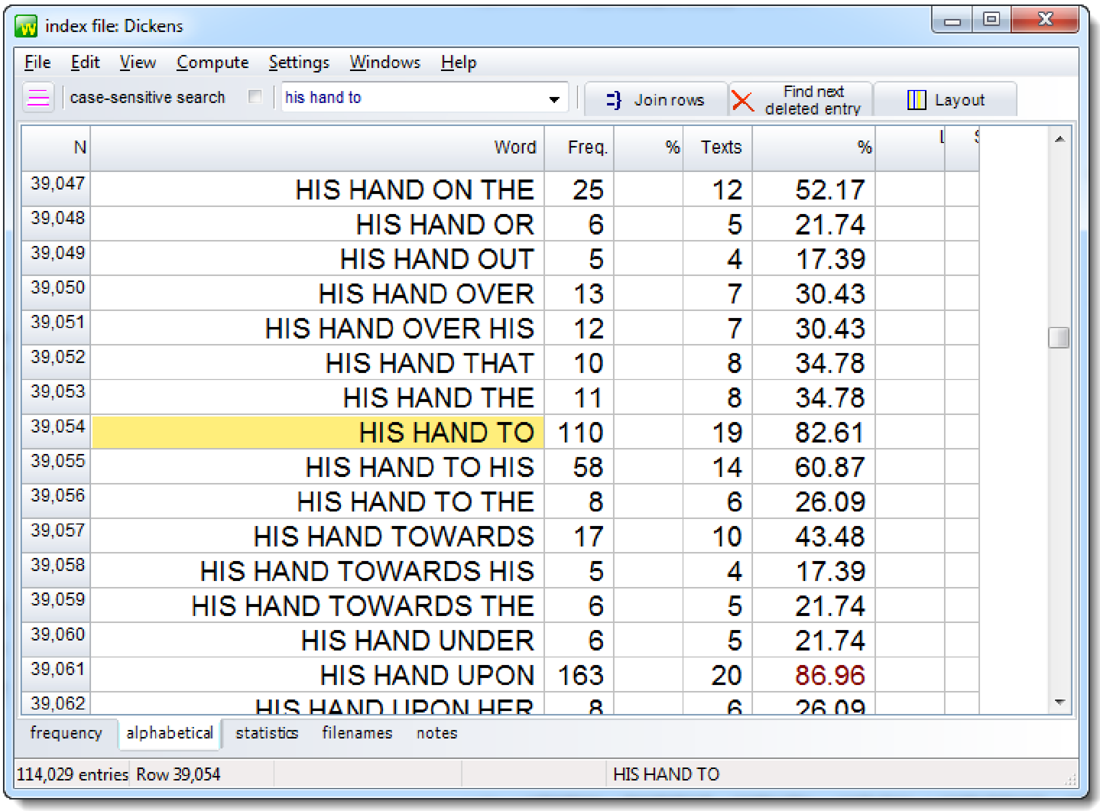
min. frequency 5

☑ omit any containing #

□ omit phrase frames

Working constraints

Here below, the listing has all **`his hand`** sequences together but not **`drawing his hand across`**, **`gave his hand to`**, etc. as shown in the phrase frame view above.

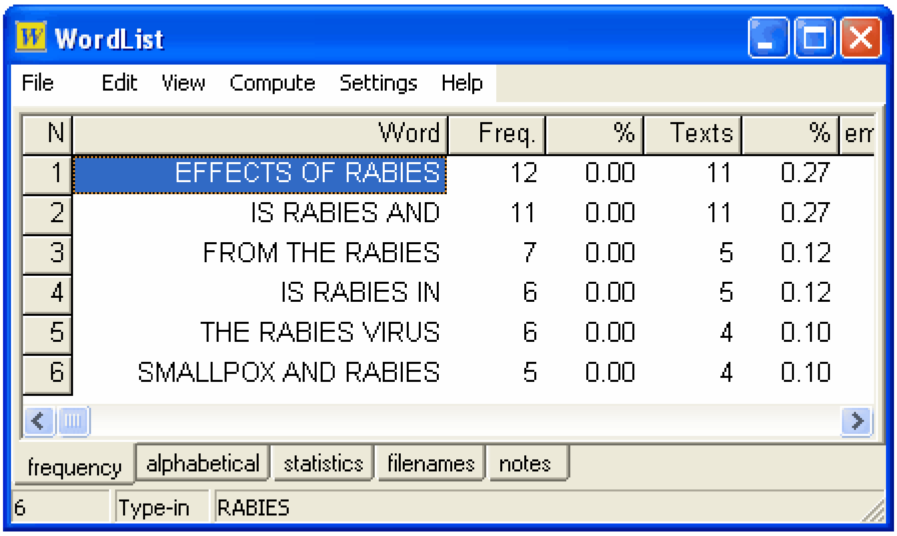


index file: Dickens
File Edit View Compute Settings Windows Help
case-sensitive search
his hand to
Join rows
Find next deleted entry
Layout
frequency alphabetical statistics filenames notes

N	Word	Freq.	%	Texts	%
39,047	HIS HAND ON THE	25		12	52.17
39,048	HIS HAND OR	6		5	21.74
39,049	HIS HAND OUT	5		4	17.39
39,050	HIS HAND OVER	13		7	30.43
39,051	HIS HAND OVER HIS	12		7	30.43
39,052	HIS HAND THAT	10		8	34.78
39,053	HIS HAND THE	11		8	34.78
39,054	HIS HAND TO	110		19	82.61
39,055	HIS HAND TO HIS	58		14	60.87
39,056	HIS HAND TO THE	8		6	26.09
39,057	HIS HAND TOWARDS	17		10	43.48
39,058	HIS HAND TOWARDS HIS	5		4	17.39
39,059	HIS HAND TOWARDS THE	6		5	21.74
39,060	HIS HAND UNDER	6		5	21.74
39,061	HIS HAND UPON	163		20	86.96
39,062	HIS HAND UPON HER	8		6	26.09

114,029 entries Row 39,054 HIS HAND TO

Here is a small set of 3-word clusters involving rabies from the BNC corpus.


WordList
File Edit View Compute Settings Help
frequency alphabetical statistics filenames notes
6 Type-in RABIES

N	Word	Freq.	%	Texts	%	em
1	EFFECTS OF RABIES	12	0.00	11	0.27	
2	IS RABIES AND	11	0.00	11	0.27	
3	FROM THE RABIES	7	0.00	5	0.12	
4	IS RABIES IN	6	0.00	5	0.12	
5	THE RABIES VIRUS	6	0.00	4	0.10	
6	SMALLPOX AND RABIES	5	0.00	4	0.10	

Some of them are plausible multi-word units.

It's a word list

Finally, remember this listing is just like a single-word word list. You can save it as a **.lst** file and open it again at any time, separately from the index.

See also: find the files for specific clusters 269, clusters in Concord 448

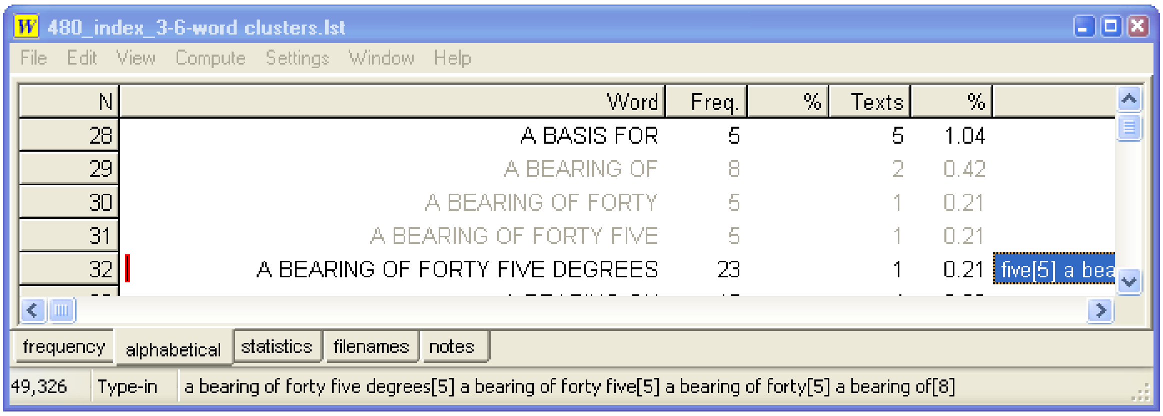
9.8.4 join clusters

The idea is to group clusters like

I DON'T THINK
NO I DON'T THINK
I DON'T THINK SO
I DON'T THINK THAT
etc.

You can join them up in a process like lemmatisation 270, either so that the smaller clusters get merged as 'lemmas' of a bigger one, or so that the smaller ones end up as 'lemmas'.

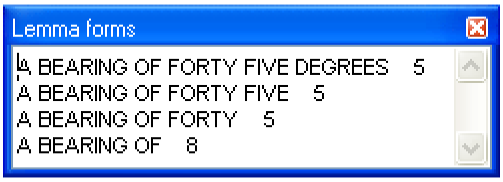
In this screenshot, shorter clusters have been merged with longer ones so that **A BEARING OF FORTY-FIVE DEGREES** relates to several related clusters:


480_index_3-6-word clusters.lst
File Edit View Compute Settings Window Help
frequency
alphabetical
statistics
filenames
notes

N	Word	Freq.	%	Texts	%	
28	A BASIS FOR	5		5	1.04	
29	A BEARING OF	8		2	0.42	
30	A BEARING OF FORTY	5		1	0.21	
31	A BEARING OF FORTY FIVE	5		1	0.21	
32	A BEARING OF FORTY FIVE DEGREES	23		1	0.21	five[5] a bea

49,326 Type-in a bearing of forty five degrees[5] a bearing of forty five[5] a bearing of forty[5] a bearing of[8]

visible by double-clicking the lemmas to show something like this:



A BEARING OF FORTY FIVE DEGREES 5
A BEARING OF FORTY FIVE 5
A BEARING OF FORTY 5
A BEARING OF 8

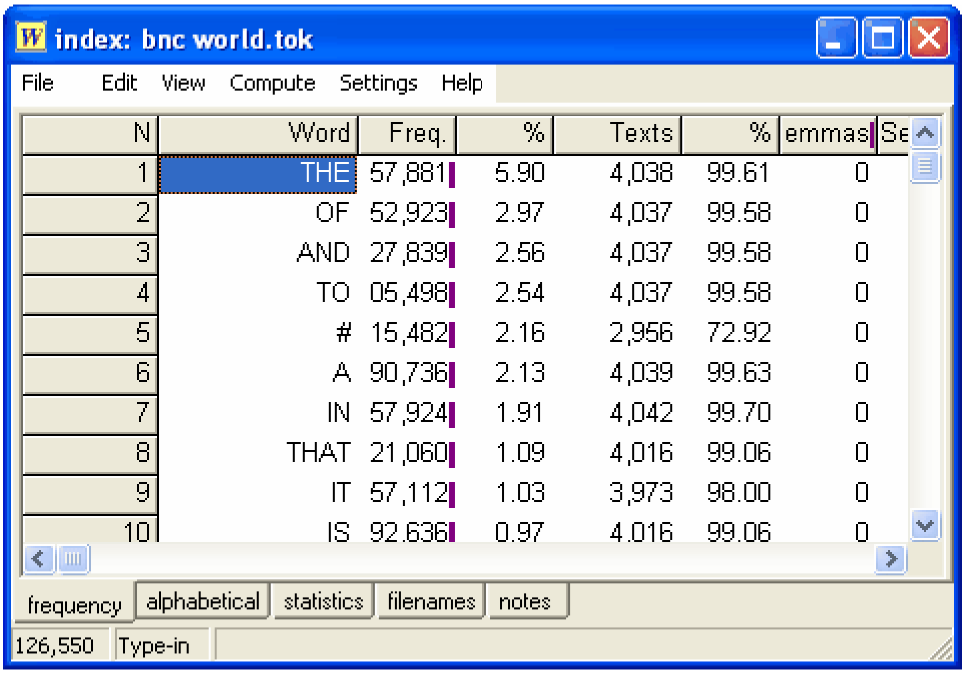
How to do it

Choose *Edit* | *Join* | *Join Clusters* in the WordList menu. The process takes quite a time because each cluster has to be compared with all those in the rest of the list; interrupt it if necessary by pressing Suspend 123.

9.8.5 index lists: viewing

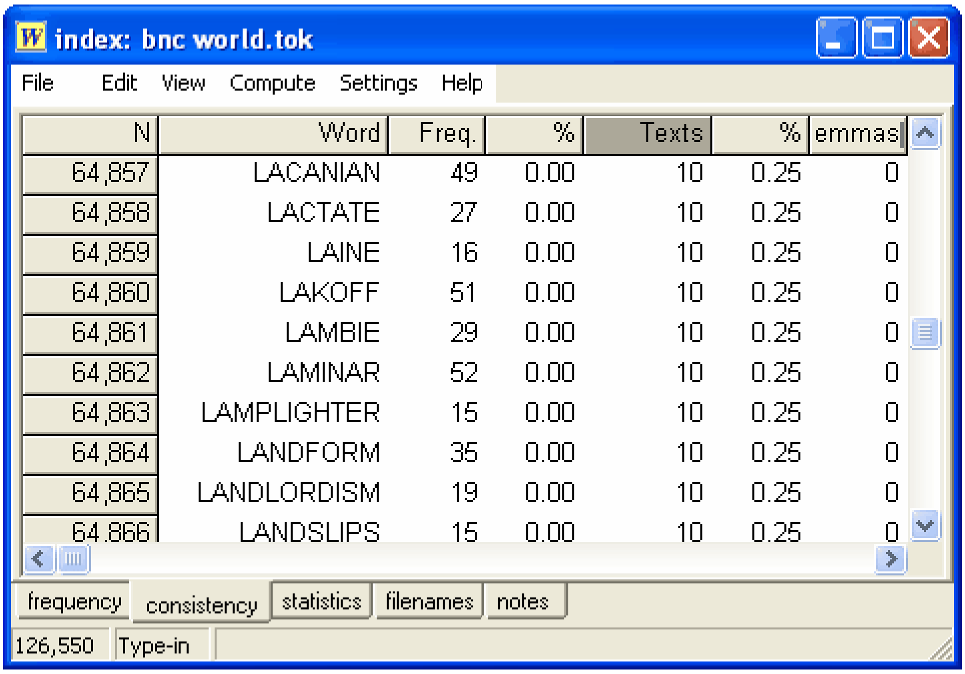
In WordList, open an index as you would any other kind of word-list file -- using File | Open. The filename will end `.tokens.` Easier, in the *Controller* | *Previous lists*, choose any index you've made and double-click it.

The index *looks* exactly like a large word-list. (Underneath, it "knows" a lot more and can do more but it looks the same.)



N	Word	Freq.	%	Texts	%	emmas	Se
1	THE	57,881	5.90	4,038	99.61	0	
2	OF	52,923	2.97	4,037	99.58	0	
3	AND	27,839	2.56	4,037	99.58	0	
4	TO	05,498	2.54	4,037	99.58	0	
5	#	15,482	2.16	2,956	72.92	0	
6	A	90,736	2.13	4,039	99.63	0	
7	IN	57,924	1.91	4,042	99.70	0	
8	THAT	21,060	1.09	4,016	99.06	0	
9	IT	57,112	1.03	3,973	98.00	0	
10	IS	92,636	0.97	4,016	99.06	0	

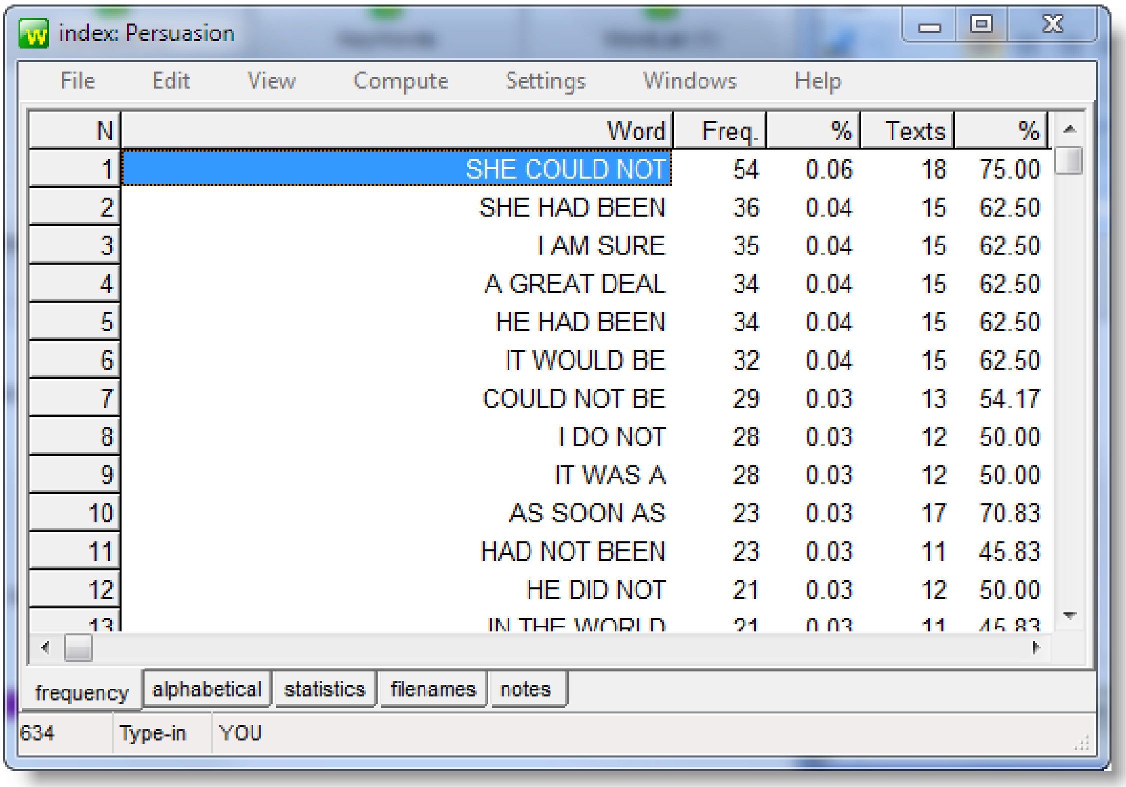
The picture above shows the top 10 words in the BNC Corpus. Number 5 (#) represents numbers or words which contain numbers such as £50.00. These very frequent words are also very consistent -- they appear in at least 99% of the 4,054 texts of BNC.
In the view below, you see words sorted by the number of Texts: all these words appeared 10 times in the corpus but their frequencies vary.



N	Word	Freq.	%	Texts	%	emmas
64,857	LACANIAN	49	0.00	10	0.25	0
64,858	LACTATE	27	0.00	10	0.25	0
64,859	LAINE	16	0.00	10	0.25	0
64,860	LAKOFF	51	0.00	10	0.25	0
64,861	LAMBIE	29	0.00	10	0.25	0
64,862	LAMINAR	52	0.00	10	0.25	0
64,863	LAMPLIGHTER	15	0.00	10	0.25	0
64,864	LANDFORM	35	0.00	10	0.25	0
64,865	LANDLORDISM	19	0.00	10	0.25	0
64,866	LANDSLIPS	15	0.00	10	0.25	0

You can highlight one or more words or mark them with the option, then C to get a speedy concordance.

But its best use to start with is to generate word clusters 278 like these:



N	Word	Freq.	%	Texts	%
1	SHE COULD NOT	54	0.06	18	75.00
2	SHE HAD BEEN	36	0.04	15	62.50
3	I AM SURE	35	0.04	15	62.50
4	A GREAT DEAL	34	0.04	15	62.50
5	HE HAD BEEN	34	0.04	15	62.50
6	IT WOULD BE	32	0.04	15	62.50
7	COULD NOT BE	29	0.03	13	54.17
8	I DO NOT	28	0.03	12	50.00
9	IT WAS A	28	0.03	12	50.00
10	AS SOON AS	23	0.03	17	70.83
11	HAD NOT BEEN	23	0.03	11	45.83
12	HE DID NOT	21	0.03	12	50.00
13	IN THE WORLD	21	0.03	11	45.83

See also Making an Index List 276, WordList clusters 278, WordList Help Contents 258.

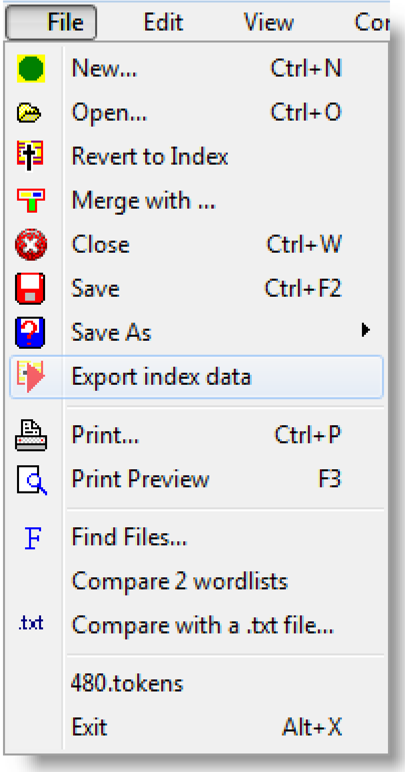
9.8.6 index exporting

The point of it...

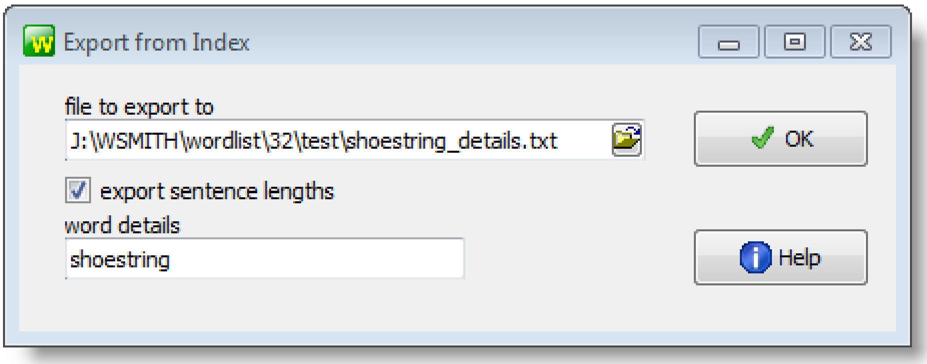
An index file knows the position of every single word in your corpus and it is possible therefore to ask it to supply specific data. For example, the lengths of each sentence or each text in the corpus (in words), or the position of each occurrence of a given word.

How to do it

With an index open, choose File | Export index data,



then complete the form with what you need.



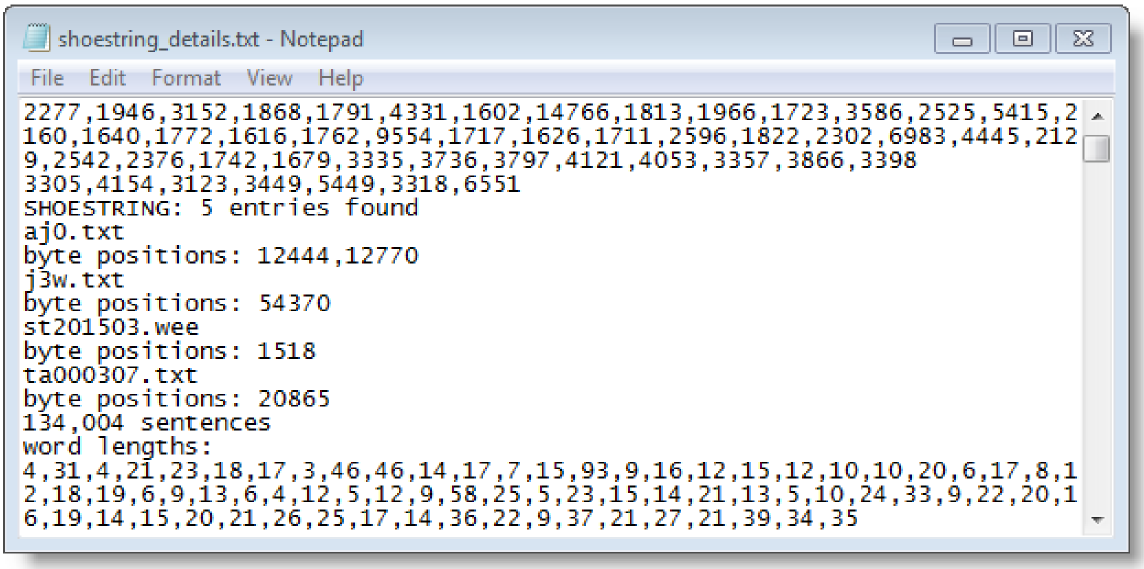
file to export to
J:\WSMITH\wordlist\32\test\shoestring_details.txt

☑ export sentence lengths

word details
shoestring

Here we have chosen to export the details about the word **SHOESTRING** in a given index, and to get to see all the sentence lengths (of all sentences in the corpus, not just the ones containing that word).

A fragment of the results are shown here:



```
2277,1946,3152,1868,1791,4331,1602,14766,1813,1966,1723,3586,2525,5415,2
160,1640,1772,1616,1762,9554,1717,1626,1711,2596,1822,2302,6983,4445,212
9,2542,2376,1742,1679,3335,3736,3797,4121,4053,3357,3866,3398
3305,4154,3123,3449,5449,3318,6551
SHOESTRING: 5 entries found
aj0.txt
byte positions: 12444,12770
j3w.txt
byte positions: 54370
st201503.wee
byte positions: 1518
ta000307.txt
byte positions: 20865
134,004 sentences
word lengths:
4,31,4,21,23,18,17,3,46,46,14,17,7,15,93,9,16,12,15,12,10,10,20,6,17,8,1
2,18,19,6,9,13,6,4,12,5,12,9,58,25,5,23,15,14,21,13,5,10,24,33,9,22,20,1
6,19,14,15,20,21,26,25,17,14,36,22,9,37,21,27,21,39,34,35
```

At the top there are word-lengths of some of the 480 text files, the last of which was 6551 words long; then we see the details of 5 cases of the word **SHOESTRING** in the corpus, which appeared twice in text AJ0.txt, once in J3W.txt etc.; finally we get the word-lengths of all the sentences in the corpus : the first one only 4 words long.

This process will be quite slow if you request a lot of data. If you don't check the sentence lengths you will still get text lengths; it wil be quicker if you leave the word details space empty.

9.9 menu search

Using the menu you can search for a sub-string within an entry -- e.g. all words containing "fore" (by entering ***fore*** -- the asterisk means that the item can be found in the middle of a word, so ***fore** will find *before* but not *beforehand*, while ***fore*** will find them both). These searches can be repeated.
This function enables you to find parts of words so that you can edit your word-list, e.g. by joining two words as one.
You can search for ends or middles of words by using the ***** wildcard.
Thus ***TH*** will find *other, something*, etc.
***TH** will find *booth, sooth*, etc.
You can then use **F8** to repeat your last search.

The search hot keys are:

F8 repeat last search (use in conjunction with F10 or F11)	
F10	search forwards from the current line
F11	search backwards from the current line
F12	search starting from the beginning

This function is handy for lemmatization 270 (joining words which belong under one entry, such as *seem/ seems/ seemed/ seeming* etc.)

See also: searching for an entry by typing 109

9.10 relationships between words

9.10.1 mutual information and other relations

the point of it

A Mutual Information (MI) score relates one word to another. For example, if *problem* is often found with *solve*, they may have a high mutual information score. Usually, *the* will be found much more often near *problem* than *solve*, so the procedure for calculating Mutual Information takes into account not just the most frequent words found near the word in question, but also whether each word is often found elsewhere, well away from the word in question. Since *the* is found very often indeed far away from *problem*, it will not tend to be related, that is, it will get a low MI score.

There are several other alternative statistics: you can see examples of how they differ here 289.

This relationship is bi-lateral: in the case of *kith* and *kin*, it doesn't distinguish between the virtual certainty of finding *kin* near *kith*, and the much lower likelihood of finding *kith* near *kin*.

There are various different formulae for computing the strength of collocational relationships. The MI in WordSmith ("specific mutual information") is computed using a formula derived from Gaussier, Lange and Meunier described in Oakes 417, p. 174; here the probability is based on total corpus size in tokens. Other measures of collocational relation are computed too, which you will see explained under Mutual Information Display 289.

Settings

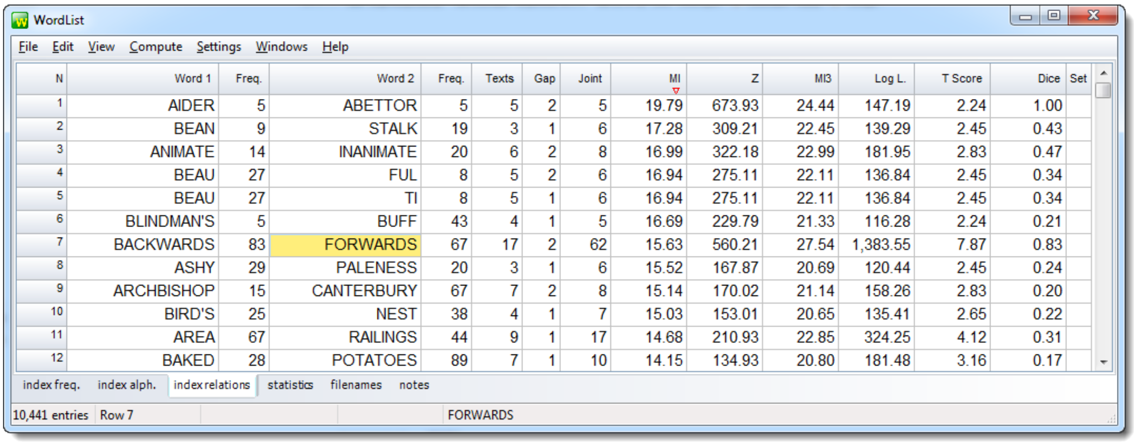
The Relationships settings are found in the Controller 4 under *Main Settings* | *Advanced* | *Index* 310 or in a menu option in **WordList**.

See also: Mutual Information Display 289, Computing Mutual Information 294, Making an Index List 276, Viewing Index Lists 284, WordList Help Contents 258.

See Oakes 417 for further information about Mutual Information, Dice, MI3 etc.

9.10.2 relationships display

The Relationships procedure contains a number of columns and uses various formulae 433:



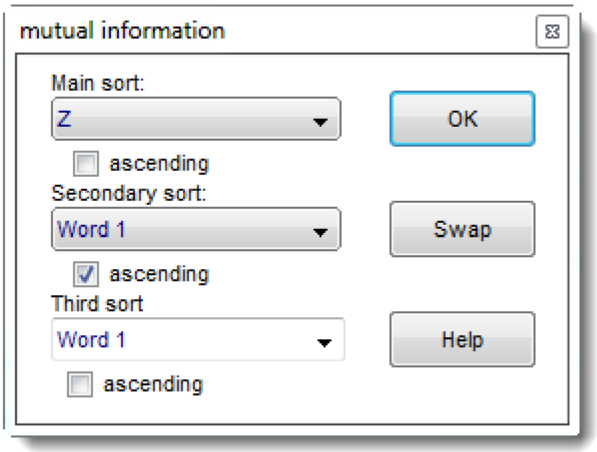
N	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI	Z	MI3	Log L.	T Score	Dice	Set
1	AIDER	5	ABETTOR	5	5	2	5	19.79	673.93	24.44	147.19	2.24	1.00	
2	BEAN	9	STALK	19	3	1	6	17.28	309.21	22.45	139.29	2.45	0.43	
3	ANIMATE	14	INANIMATE	20	6	2	8	16.99	322.18	22.99	181.95	2.83	0.47	
4	BEAU	27	FUL	8	5	2	6	16.94	275.11	22.11	136.84	2.45	0.34	
5	BEAU	27	TI	8	5	1	6	16.94	275.11	22.11	136.84	2.45	0.34	
6	BLINDMAN'S	5	BUFF	43	4	1	5	16.69	229.79	21.33	116.28	2.24	0.21	
7	BACKWARDS	83	FORWARDS	67	17	2	62	15.63	560.21	27.54	1,383.55	7.87	0.83	
8	ASHY	29	PALENESS	20	3	1	6	15.52	167.87	20.69	120.44	2.45	0.24	
9	ARCHBISHOP	15	CANTERBURY	67	7	2	8	15.14	170.02	21.14	158.26	2.83	0.20	
10	BIRD'S	25	NEST	38	4	1	7	15.03	153.01	20.65	135.41	2.65	0.22	
11	AREA	67	RAILINGS	44	9	1	17	14.68	210.93	22.85	324.25	4.12	0.31	
12	BAKED	28	POTATOES	89	7	1	10	14.15	134.93	20.80	181.48	3.16	0.17	

10,441 entries Row 7 FORWARDS

Word 1: the first word in a pair, followed by Freq. (its frequency in the whole index).
Word 2: the other word in that pair, followed by Freq. (its frequency in the whole index). If you have computed "to right only 295", then Word 1 precedes Word 2.
Texts: the number of texts this pair was found in (there were 23 in the whole index).
Gap: the most typical distance between Word 1 and Word 2.
Joint: their joint frequency over the entire span 294 (not just the joint frequency at the typical gap distance).

In line 7 of this display, **BACKWARDS** occurs 83 times in the whole index (based on Dickens novels), and **FORWARDS** 8 times. They occur together 62 times. The gap is 2 because **backwards**, in these data, typically comes 2 words away from **forwards**. The pair **backwards * forwards** comes in 17 texts. (This search was computed using the to right only 295 setting mentioned above).

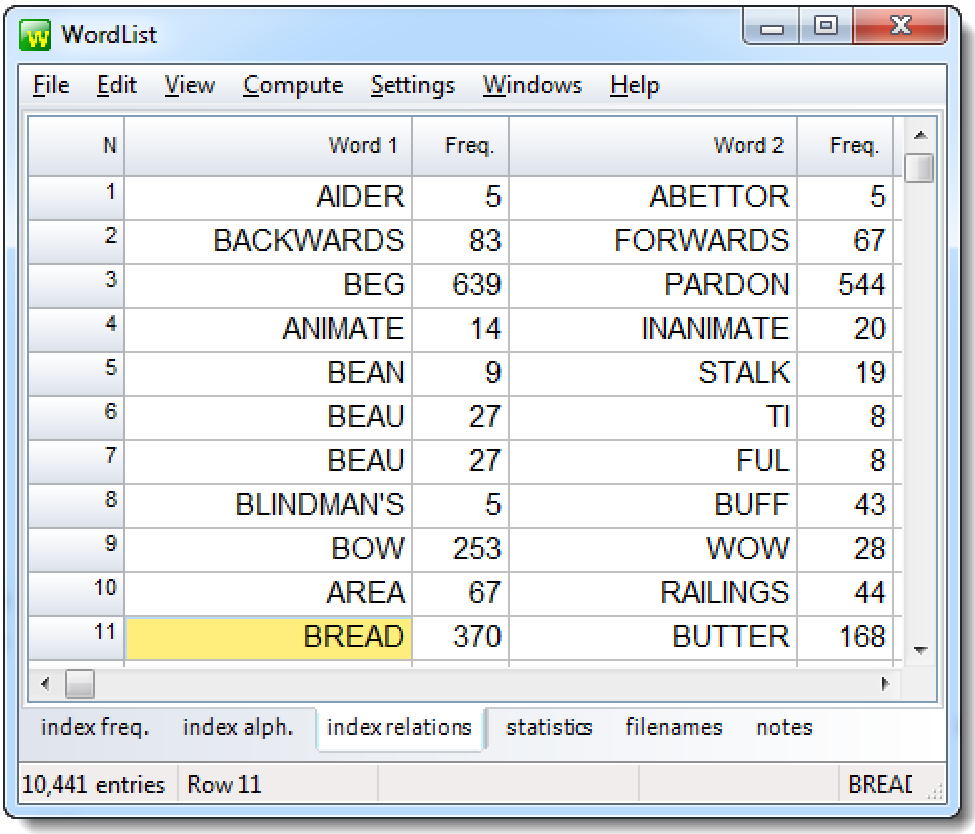
As usual, the data can be sorted by clicking on the headers. Let's now sort by clicking on "Z score" first and "Word 1" second.



mutual information

Main sort: Z
□ ascending
Secondary sort: Word 1
☑ ascending
Third sort Word 1
□ ascending

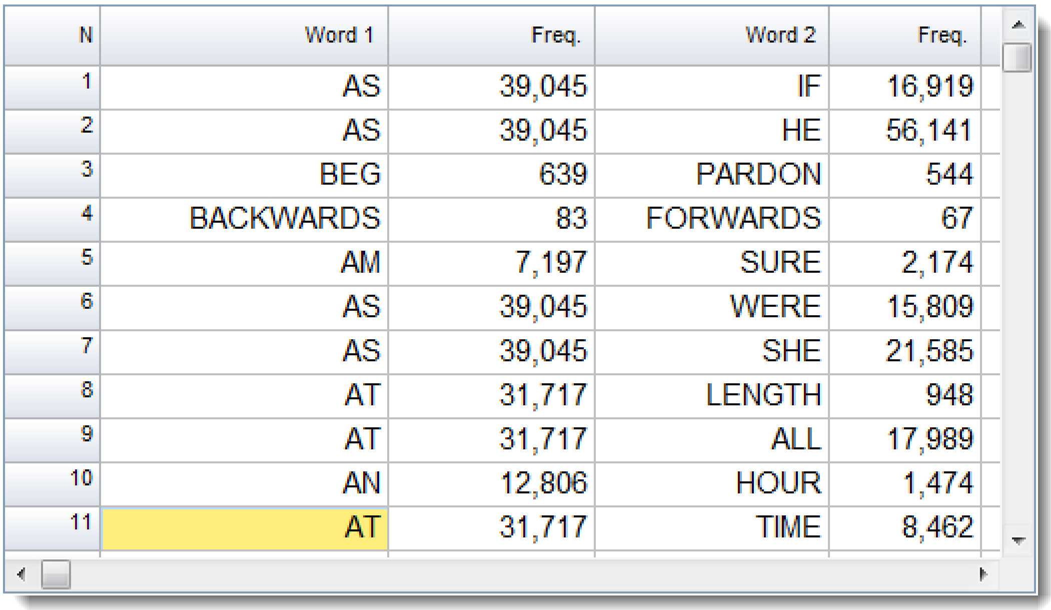
You get a double sort, main and secondary, because sometimes you will want to see how MI or Z score or other sorting affects the whole list and sometimes you will want to keep the words sorted alphabetically and only sort by MI or Z score within each word-type. Press *Swap* to switch the primary & secondary sorts.


WordList
File Edit View Compute Settings Windows Help
index freq. index alph. index relations statistics filenames notes
10,441 entries Row 11 BREAD

N	Word 1	Freq.	Word 2	Freq.
1	AIDER	5	ABETTOR	5
2	BACKWARDS	83	FORWARDS	67
3	BEG	639	PARDON	544
4	ANIMATE	14	INANIMATE	20
5	BEAN	9	STALK	19
6	BEAU	27	TI	8
7	BEAU	27	FUL	8
8	BLINDMAN'S	5	BUFF	43
9	BOW	253	WOW	28
10	AREA	67	RAILINGS	44
11	BREAD	370	BUTTER	168

The order is not quite the same ... but not very different either. Both Freq. columns have fairly small numbers.

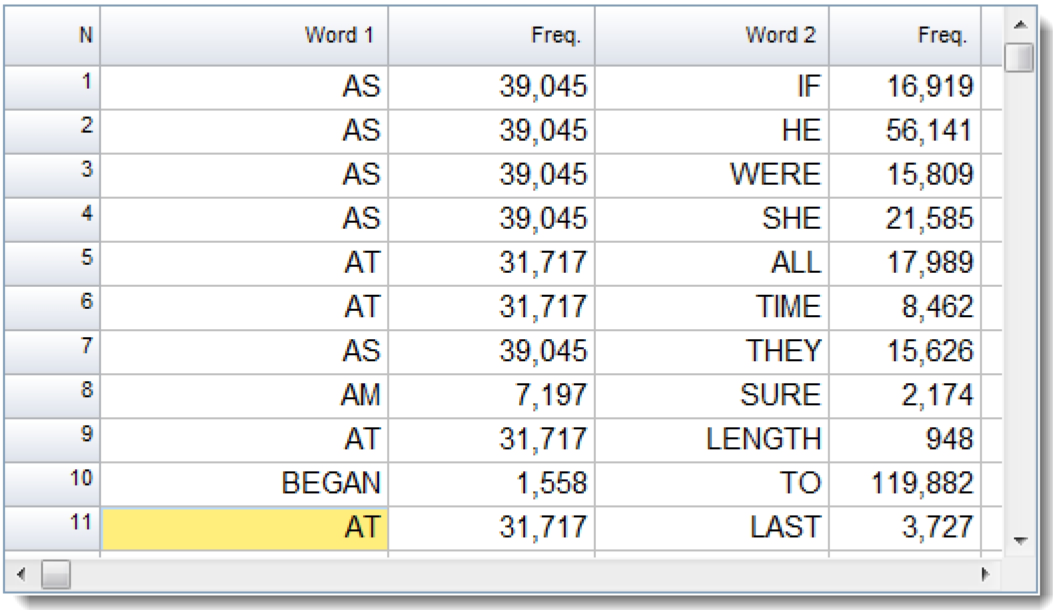
Here is the display sorted by MI3 Score (Oakes 417 p. 172):



N	Word 1	Freq.	Word 2	Freq.
1	AS	39,045	IF	16,919
2	AS	39,045	HE	56,141
3	BEG	639	PARDON	544
4	BACKWARDS	83	FORWARDS	67
5	AM	7,197	SURE	2,174
6	AS	39,045	WERE	15,809
7	AS	39,045	SHE	21,585
8	AT	31,717	LENGTH	948
9	AT	31,717	ALL	17,989
10	AN	12,806	HOUR	1,474
11	AT	31,717	TIME	8,462

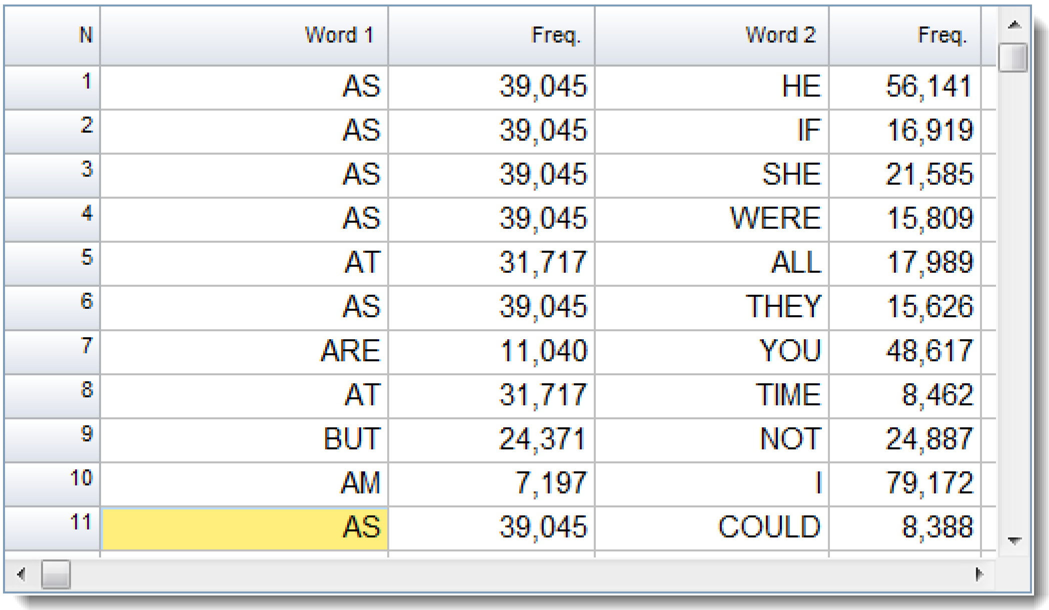
Much more frequent items have jumped to the top.

Now, by Log Likelihood (Dunning 417, 1993):



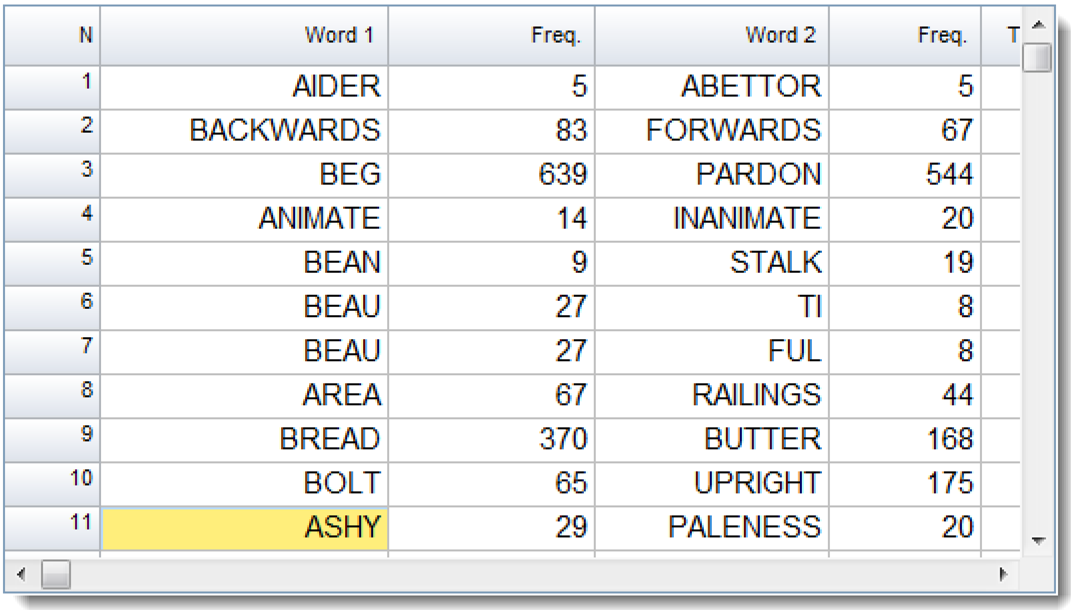
N	Word 1	Freq.	Word 2	Freq.
1	AS	39,045	IF	16,919
2	AS	39,045	HE	56,141
3	AS	39,045	WERE	15,809
4	AS	39,045	SHE	21,585
5	AT	31,717	ALL	17,989
6	AT	31,717	TIME	8,462
7	AS	39,045	THEY	15,626
8	AM	7,197	SURE	2,174
9	AT	31,717	LENGTH	948
10	BEGAN	1,558	TO	119,882
11	AT	31,717	LAST	3,727

Here the Word 2 items are again very high frequency ones and we get at colligation (grammatical collocation). A T Score listing is fairly similar:



N	Word 1	Freq.	Word 2	Freq.
1	AS	39,045	HE	56,141
2	AS	39,045	IF	16,919
3	AS	39,045	SHE	21,585
4	AS	39,045	WERE	15,809
5	AT	31,717	ALL	17,989
6	AS	39,045	THEY	15,626
7	ARE	11,040	YOU	48,617
8	AT	31,717	TIME	8,462
9	BUT	24,371	NOT	24,887
10	AM	7,197	I	79,172
11	AS	39,045	COULD	8,388

but a Dice score ordered list brings us back to results akin to the first two shown above:



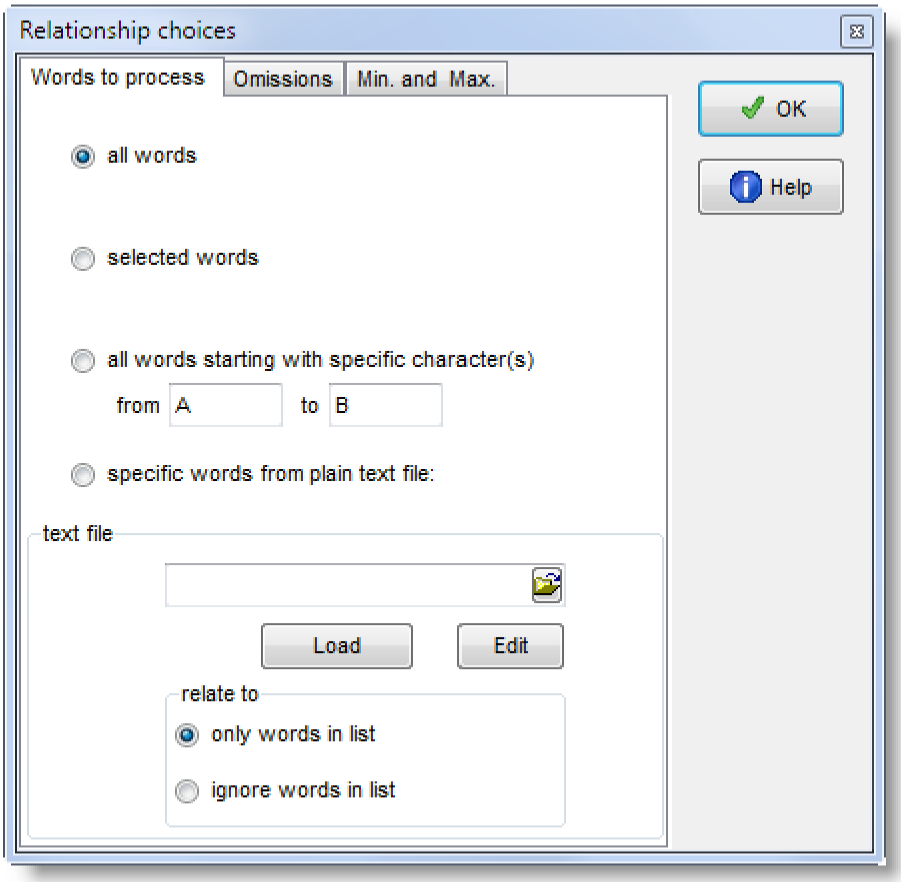
N	Word 1	Freq.	Word 2	Freq.	T
1	AIDER	5	ABETTOR	5	
2	BACKWARDS	83	FORWARDS	67	
3	BEG	639	PARDON	544	
4	ANIMATE	14	INANIMATE	20	
5	BEAN	9	STALK	19	
6	BEAU	27	TI	8	
7	BEAU	27	FUL	8	
8	AREA	67	RAILINGS	44	
9	BREAD	370	BUTTER	168	
10	BOLT	65	UPRIGHT	175	
11	ASHY	29	PALENESS	20	

See also: Formulae 433, Mutual Information and other relationships 289, Computing Relationships 294, Making an Index List 276, Viewing Index Lists 284, WordList Help Contents 258.

See Oakes 417 for further information about the various statistics offered.

9.10.3 relationships computing

To compute these relationship statistics you need a WordList Index 276. Then in its menu, choose Compute | Relationships.



Relationship choices

Words to process | Omissions | Min. and Max.

- ● all words
- ○ selected words
- ○ all words starting with specific character(s)
  from A to B
- ○ specific words from plain text file:

text file

Load | Edit

relate to
- ● only words in list
- ○ ignore words in list

words to process

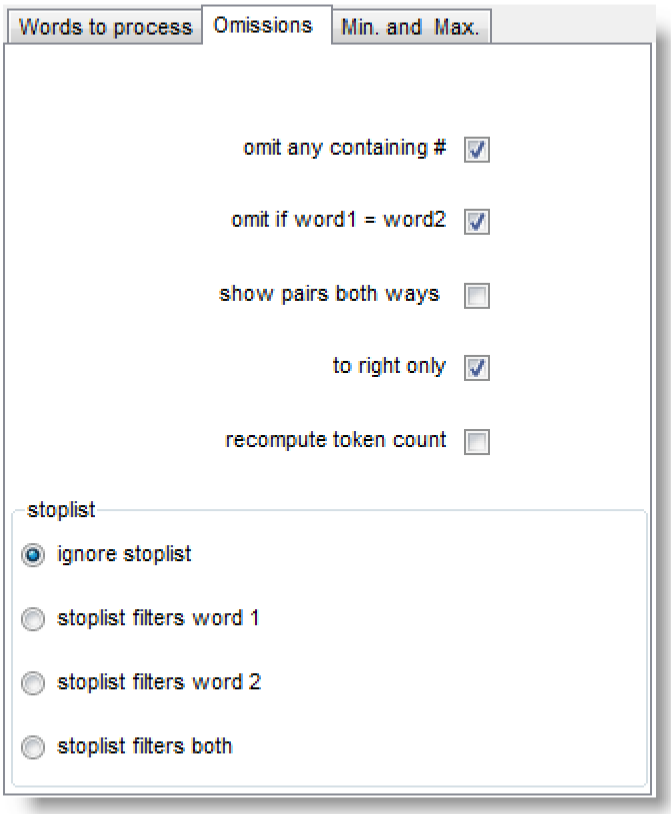
You can choose whether to compute the statistics for all entries, or only any selected (highlighted) entries, or only those between two initial characters e.g. between A and D, or indeed to use your own specified words only.

If you wish to select only a few items for MI calculation, you can mark them first 111 (with). Or you can always do part of the list (eg. A to D) and later merge 262 your mutual-information list with another (E to H).

Alternatively you may choose to use only items from a plain text file constructed using the same syntax as a match-list file., or to use all items except ones from your plain text file.

omissions

omit any containing # will cut out numbers, and *omit if word1=word2* is there because you might find that **GOOD** is related to **GOOD** if there are lots of cases where these 2 are found near each other.



omit any containing # ☑

omit if word1 = word2 ☑

show pairs both ways □

to right only ☑

recompute token count □

stoplist

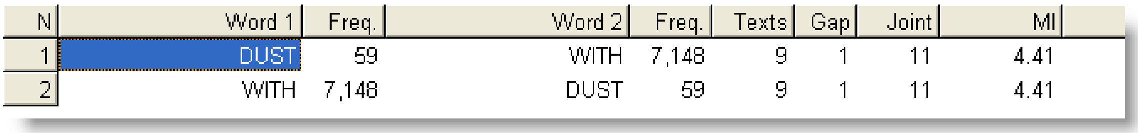
● ignore stoplist

○ stoplist filters word 1

○ stoplist filters word 2

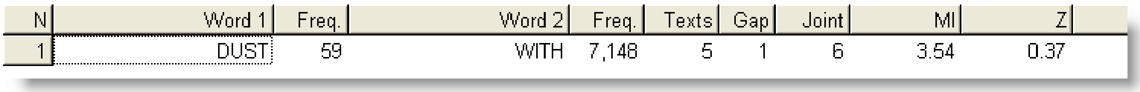
○ stoplist filters both

show pairs both ways allows you to locate all the pairs more easily because it doubles up the list. For example, suppose we have a pair of words such as **HEAVEN** and **EARTH**. This will normally enter the list only in one order, let us say **HEAVEN** as word 1 and **EARTH** as word 2. If you're looking at all the words in the Word 1 column, you will not find **EARTH**. If you want to be able to see the pair as both **HEAVEN** - **EARTH** and **EARTH** - **HEAVEN**, select *show pairs both ways*. Here we can see this with **DUST** and **WITH**



N	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI
1	DUST	59	WITH	7,148	9	1	11	4.41
2	WITH	7,148	DUST	59	9	1	11	4.41

to right only: if this is checked, possible relations are computed to the right of the node only. That is, when considering **DUST**, say, cases of **WITH** to the right will be noticed but cases where **WITH** is to the left of **DUST** would get ignored.

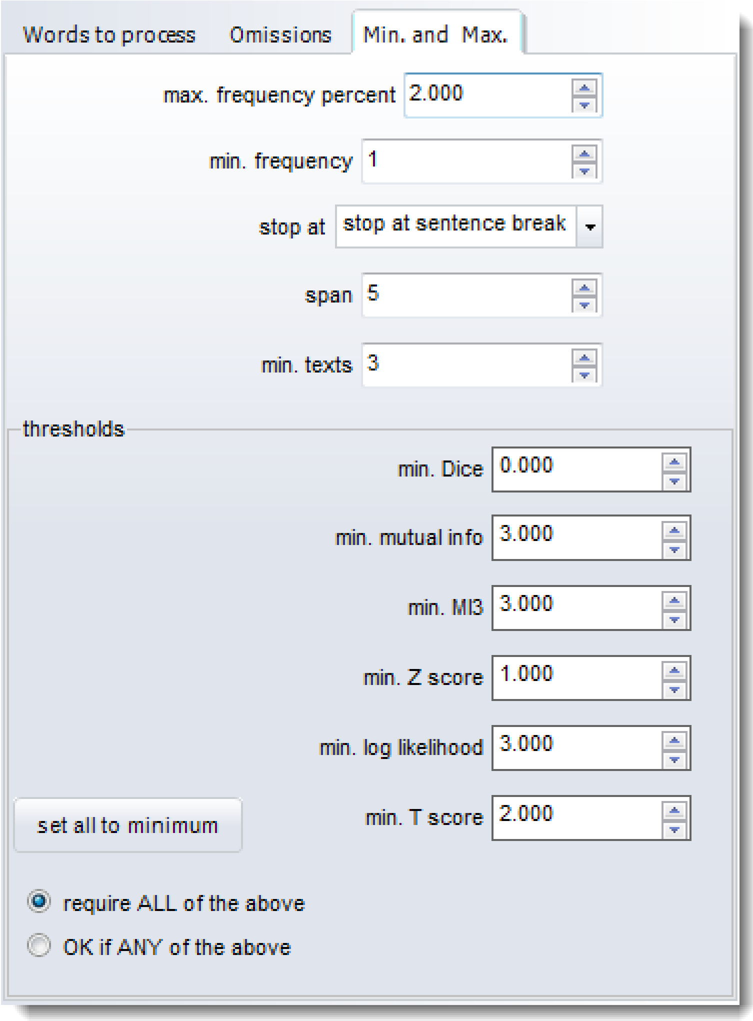


N	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI	Z
1	DUST	59	WITH	7,148	5	1	6	3.54	0.37

Here, the number of texts goes down to 5 from 9, MI score is lower, etc, because the process looks only to the right. (In the case of a right-to-left language like Arabic, the processing is still of the words following the node word.)

recompute token 297 *count* allows you to get the number of tokens counted again e.g. after items have been edited or deleted.

min. and max



max. frequency percent 2.000

min. frequency 1

stop at stop at sentence break

span 5

min. texts 3

thresholds

min. Dice 0.000

min. mutual info 3.000

min. MI3 3.000

min. Z score 1.000

min. log likelihood 3.000

min. T score 2.000

● require ALL of the above

○ OK if ANY of the above

max. frequency percent: ignores any tokens which are more frequent than the percentage indicated. Set the maximum frequency, for example, to 0.5% to cut out words whose frequency is greater than that.(The point of this is to avoid computing mutual information for words like **the** and **of**, which are likely to have a frequency greater than say 1.0%. For example 0.5%, in the case of the BNC, would mean ignoring about 20 of the top frequency words, such as **WITH, HE, YOU**. 0.1% would cut about 100 words including **GET, BACK, BECAUSE**. If you want to include all words, then set this to 100.000)

min. frequency: the minimum frequency for any item to be considered for the calculation. (Default = 5; a minimum frequency of 5 means that no word of frequency 4 or less in the index will be visible in the relationship results. If an item occurs only once or twice, the relationship is unlikely to be informative.)

stop at allows you to ignore potential relationships e.g. across sentence boundaries. It has to do with whether breaks such as punctuation or sentence breaks 188 determine that one word cannot be related to another. With stop at sentence break, "**I wrote the letter.**

`Then I posted it`" would not consider **`posted`** as a possible collocate of **`letter`** because there's a sentence break between them.

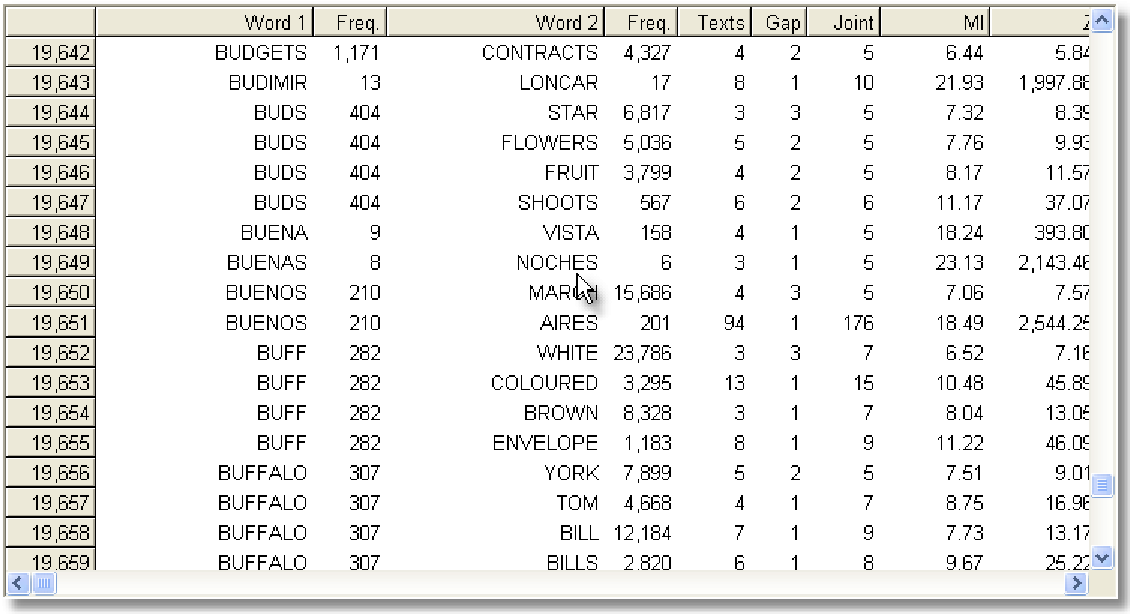
span: the number of intervening words between collocate and node. With a span of 5, the node **`wrote`** would consider **`the, letter, then, I`** and **`posted`** as possible collocates if *stop at* were set at *no limits* in the example above.

min. texts: the minimum number of texts any item must be found in to be considered for the calculation.

min. Dice/mutual info.MI3 etc: the minimum number which the MI or other selected statistic must come up with to be reported. A useful limit for MI is 3.0. Below this, the linkage between node and collocate is likely to be rather tenuous.

Choose whether ALL the values set here are used when deciding whether to show a possible relationship or ANY. (Each threshold can be set between -9999.0 and 9999.0.)

Computing the MI score for each and every entry in an index takes a long time: some years ago it took over an hour to compute MI for all words beginning with B in the case of the BNC edition (written, 90 million words) in the screenshot below, using default settings. It might take 24 hours to process the whole BNC, 100 million words, even on a modern powerful PC. Don't forget to save your results afterwards!



	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI	Z
19,642	BUDGETS	1,171	CONTRACTS	4,327	4	2	5	6.44	5.84
19,643	BUDIMIR	13	LONCAR	17	8	1	10	21.93	1,997.88
19,644	BUDS	404	STAR	6,817	3	3	5	7.32	8.39
19,645	BUDS	404	FLOWERS	5,036	5	2	5	7.76	9.93
19,646	BUDS	404	FRUIT	3,799	4	2	5	8.17	11.57
19,647	BUDS	404	SHOOTS	567	6	2	6	11.17	37.07
19,648	BUENA	9	VISTA	158	4	1	5	18.24	393.80
19,649	BUENAS	8	NOCHES	6	3	1	5	23.13	2,143.46
19,650	BUENOS	210	MARCH	15,686	4	3	5	7.06	7.57
19,651	BUENOS	210	AIRES	201	94	1	176	18.49	2,544.25
19,652	BUFF	282	WHITE	23,786	3	3	7	6.52	7.16
19,653	BUFF	282	COLOURED	3,295	13	1	15	10.48	45.89
19,654	BUFF	282	BROWN	8,328	3	1	7	8.04	13.05
19,655	BUFF	282	ENVELOPE	1,183	8	1	9	11.22	46.09
19,656	BUFFALO	307	YORK	7,899	5	2	5	7.51	9.01
19,657	BUFFALO	307	TOM	4,668	4	1	7	8.75	16.96
19,658	BUFFALO	307	BILL	12,184	7	1	9	7.73	13.17
19,659	BUFFALO	307	BILLS	2,820	6	1	8	9.67	25.22

See also Collocates 179, Mutual Information Settings 289, Mutual Information Display 289, Detailed Consistency Relations 268, Making an Index List 276, Viewing Index Lists 284, Recompute Token Count 297, WordList Help Contents 258.

9.11 recompute tokens

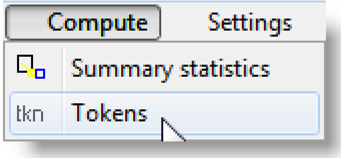
Why recompute the tokens?

To compute relations such as Mutual Information 289 or Keyness 245 we need an estimate of the total number of running words (let's call it TNR) in the text corpus from which the data came. It is tricky to decide what actually counts as the TNR. Not only are there problems to do with hyphenation 125, apostrophes and other non-letters 125 in the middle of a word, numbers 125, words cut out because of a stoplist 120 etc, but also a decision whether TNR should in principle include all of those or in principle include only the words or clusters now in the list in question. In practice for single-word word lists this usually makes little difference. In the case of word clusters, however, there might be a big difference between the TNR words and TNR clusters, and anyway what exactly is meant by running clusters of words if you think about how they are computed 448?

For most normal purposes, the total number of running words (tokens 426) computed when the word list or index was created will be used for these statistical calculations.

How to do it

Compute | Tokens



What it affects

Any decision made here will apply equally both to the node and the collocate whether these are clusters or single words, or to the little word-list and the reference corpus word-list in the case of key words calculations.

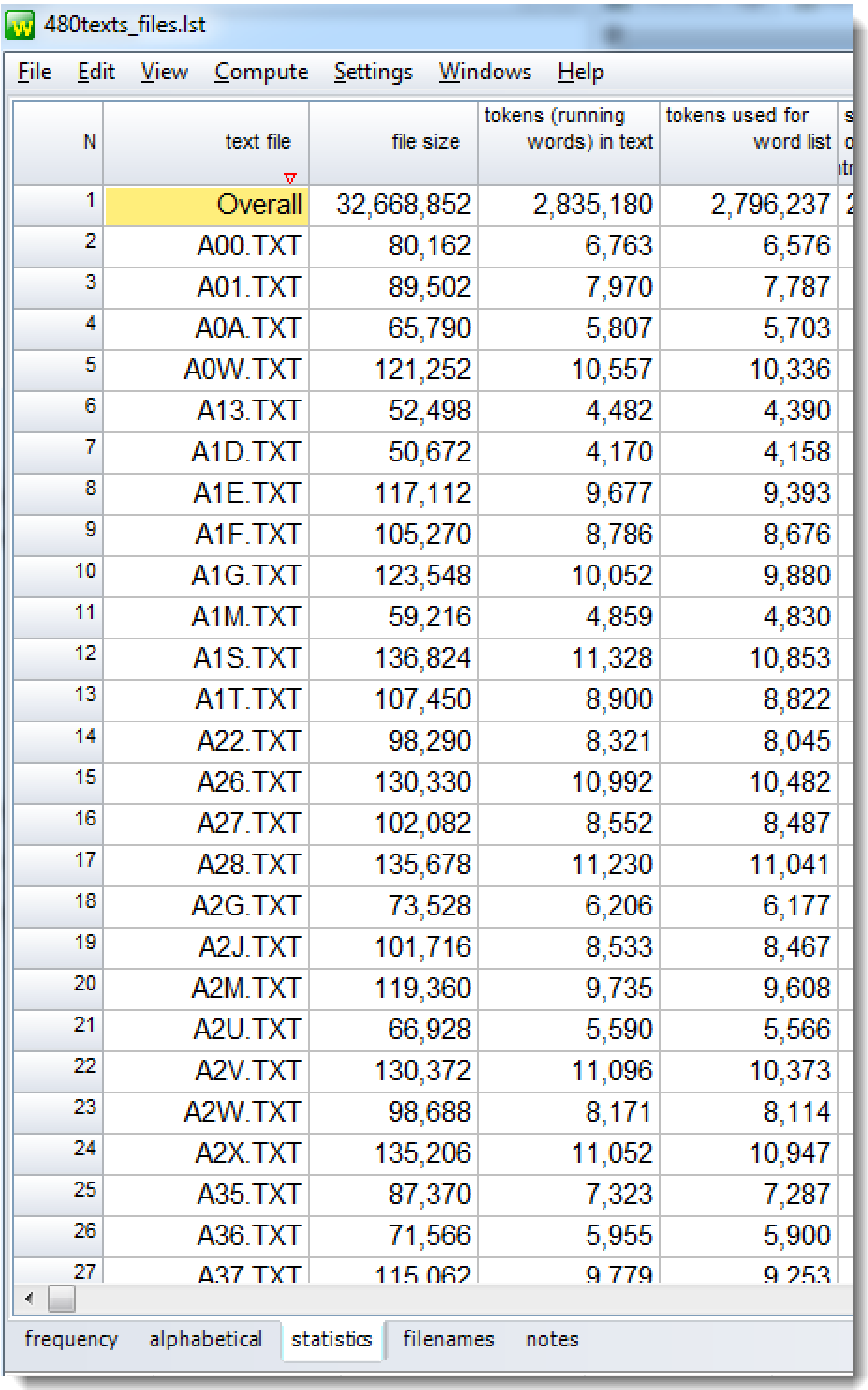
If you do choose to recompute the token count, then the TNR will be calculated as the total of the word or cluster frequencies for those entries still left in the list. After any have been zapped or if a minimum frequency above 1 is used the difference may be quite large.

If you choose *not* to recompute, the total number of running words (tokens) computed when the word list or index was created will be used.

9.12 statistics

9.12.1 statistics

Visible by clicking the Statistics tab at the bottom of a **WordList** window:


480texts_files.lst
File Edit View Compute Settings Windows Help
frequency alphabetical statistics filenames notes

N	text file	file size	tokens (running words) in text	tokens used for word list
1	Overall	32,668,852	2,835,180	2,796,237
2	A00.TXT	80,162	6,763	6,576
3	A01.TXT	89,502	7,970	7,787
4	A0A.TXT	65,790	5,807	5,703
5	A0W.TXT	121,252	10,557	10,336
6	A13.TXT	52,498	4,482	4,390
7	A1D.TXT	50,672	4,170	4,158
8	A1E.TXT	117,112	9,677	9,393
9	A1F.TXT	105,270	8,786	8,676
10	A1G.TXT	123,548	10,052	9,880
11	A1M.TXT	59,216	4,859	4,830
12	A1S.TXT	136,824	11,328	10,853
13	A1T.TXT	107,450	8,900	8,822
14	A22.TXT	98,290	8,321	8,045
15	A26.TXT	130,330	10,992	10,482
16	A27.TXT	102,082	8,552	8,487
17	A28.TXT	135,678	11,230	11,041
18	A2G.TXT	73,528	6,206	6,177
19	A2J.TXT	101,716	8,533	8,467
20	A2M.TXT	119,360	9,735	9,608
21	A2U.TXT	66,928	5,590	5,566
22	A2V.TXT	130,372	11,096	10,373
23	A2W.TXT	98,688	8,171	8,114
24	A2X.TXT	135,206	11,052	10,947
25	A35.TXT	87,370	7,323	7,287
26	A36.TXT	71,566	5,955	5,900
27	A37.TXT	115,062	9,779	9,253

Overall results take up the top row. Details for the individual text files follow below.

Statistics include:

number of files involved in the word-list
file size (in bytes, i.e. characters)

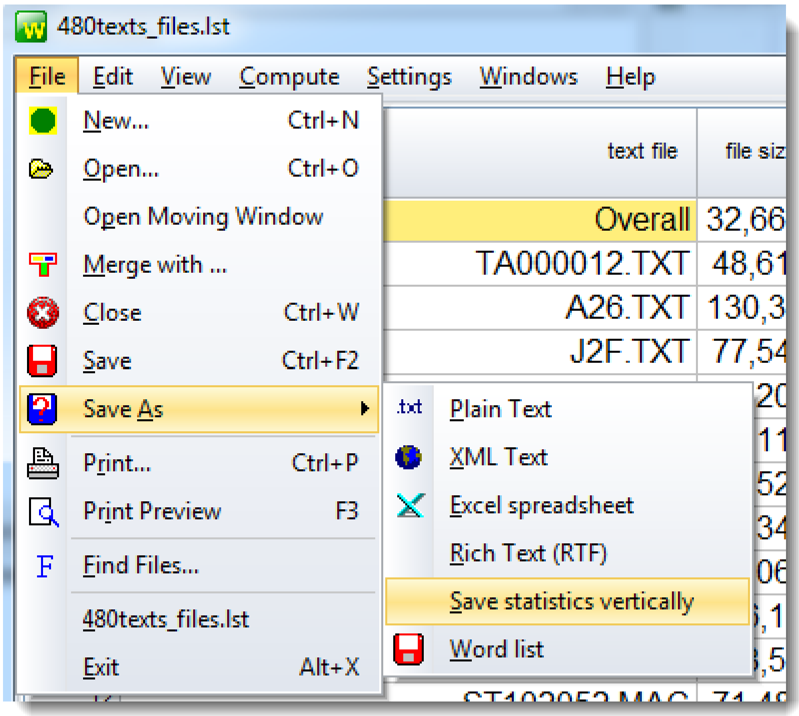
running words in the text (***tokens***)
tokens used in the list (would be affected by using a stoplist 120 or changes to minimum settings 314)
sum of entries: choose *Compute* | *Tokens* to see, otherwise this will be blank
no. of different words (***types***)
type/token ratios 303
no. of sentences 147 in the text
mean sentence length (in words)
standard deviation of sentence length (in words)
no. of paragraphs 147 in the text
mean paragraph length (in words)
standard deviation of paragraph length (in words)
no. of headings 146 in the text (none here because WordSmith didn't know how to recognise headings)
mean heading length (in words)
no. of sections 147 in the text (here 480 because WordSmith only noticed 1 section per text)
mean section length (in words)
standard deviation of heading length (in words)
numbers 125 removed
stoplist 120 tokens and types removed
the number of 1-letter words
...
the number of n-letter words (to see these scroll the grid horizontally)

(14 is the default 113 maximum word length. But you can set it to any length up to 50 letters in Word List Settings, in the Settings menu.) Longer words are cut short but this is indicated with a + at the end of the word.

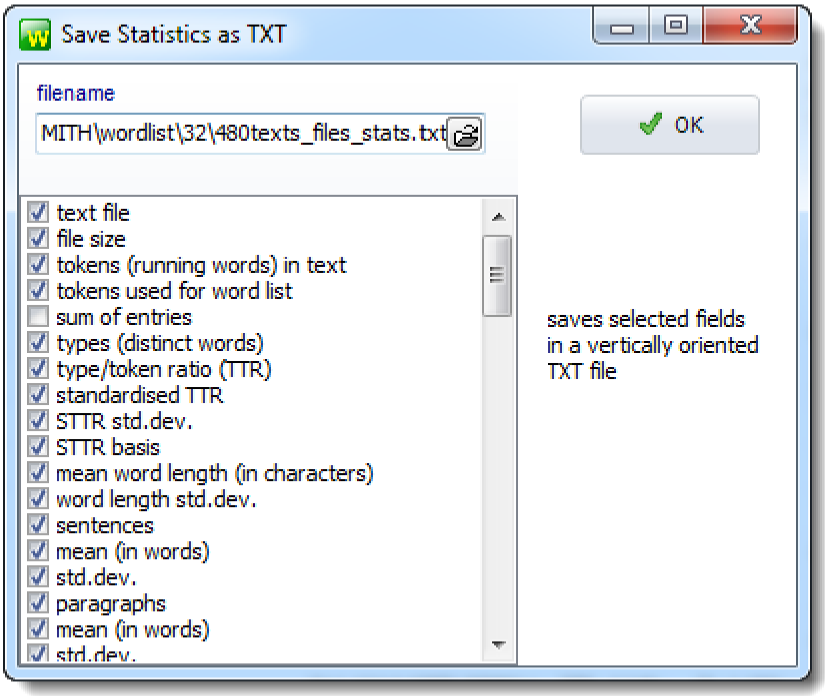
The number of types (different words) is computed separately for each text. Therefore if you have done a single word-list involving more than one text, summing the number of types for each text will not give the same total as the number of types over the whole collection.

Vertical layout

If you prefer the layout in previous versions of WordSmith, you can choose to save the statistics vertically in a text file.



This lets you choose which ones (any unchecked are zero in the data):



filename

MITH\wordlist\32\480texts_files_stats.txt

- ☑ text file
- ☑ file size
- ☑ tokens (running words) in text
- ☑ tokens used for word list
- □ sum of entries
- ☑ types (distinct words)
- ☑ type/token ratio (TTR)
- ☑ standardised TTR
- ☑ STTR std.dev.
- ☑ STTR basis
- ☑ mean word length (in characters)
- ☑ word length std.dev.
- ☑ sentences
- ☑ mean (in words)
- ☑ std.dev.
- ☑ paragraphs
- ☑ mean (in words)
- ☑ std.dev.

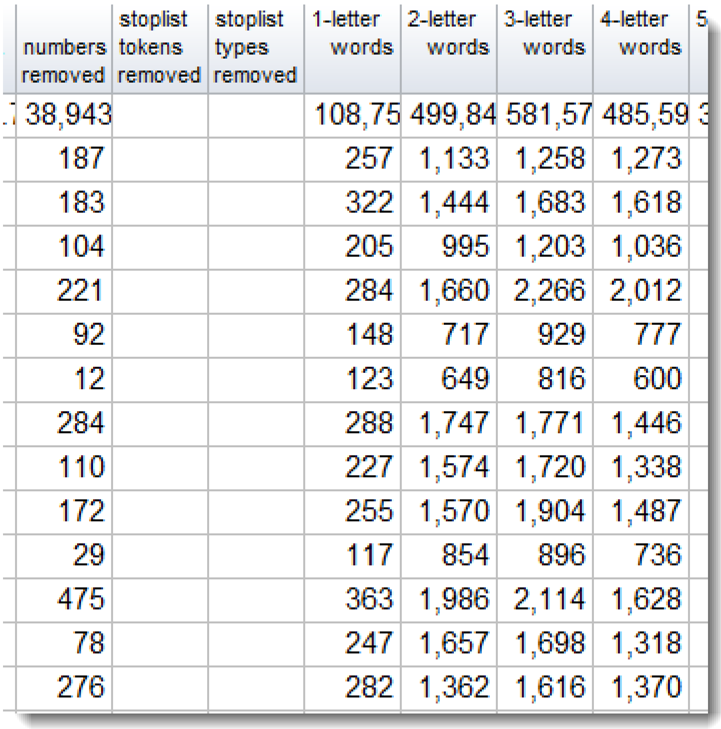
saves selected fields in a vertically oriented TXT file

and the data will be saved listed vertically.

Alternatively you could export the data here to Excel and use its Transpose function to get the rows and columns swapped.

Tokens used for word list

In these data, there were over 2.8 million running words of text, but 38,943 numbers were not listed separately, so the number of tokens in the word-list is a little under 2.8 million.

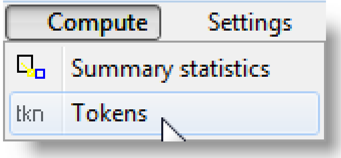


numbers removed	stoplist tokens removed	stoplist types removed	1-letter words	2-letter words	3-letter words	4-letter words	5
38,943			108,75	499,84	581,57	485,59	3
187			257	1,133	1,258	1,273	
183			322	1,444	1,683	1,618	
104			205	995	1,203	1,036	
221			284	1,660	2,266	2,012	
92			148	717	929	777	
12			123	649	816	600	
284			288	1,747	1,771	1,446	
110			227	1,574	1,720	1,338	
172			255	1,570	1,904	1,487	
29			117	854	896	736	
475			363	1,986	2,114	1,628	
78			247	1,657	1,698	1,318	
276			282	1,362	1,616	1,370	

MS Word's word count is different!

The number of tokens found is affected by your settings such as treatment of numbers 125, hyphens 435 and mid-word letter settings 125 (e.g. the apostrophe). For that reason you may well find that different programs give different values for the same text. (Besides, in the case of MS Word we are not told how a "word" is defined...)

Sum of entries can be computed after the word-list is created by choosing *Compute* | *Tokens*



and will show the total number of tokens now available by adding the **frequencies** of each entry (you may have deleted some).

See also : WordList display 318 (with a screenshot), Summary Statistics 66, Starts and Ends of Text Segments 146, Recomputing tokens 297.

9.12.2 type/token ratios

If a text is 1,000 words long, it is said to have 1,000 "tokens 426". But a lot of these words will be repeated, and there may be only say 400 different words in the text. "Types 426", therefore, are the different words.

The ratio between types and tokens in this example would be 40%.

But this type/token ratio (TTR) varies very widely in accordance with the length of the text -- or corpus of texts -- which is being studied. A 1,000 word article might have a TTR of 40%; a shorter one might reach 70%; 4 million words will probably give a type/token ratio of about 2%, and so on. Such type/token information is rather meaningless in most cases, though it is supplied in a WordList statistics display. The conventional TTR is informative, of course, if you're dealing with a corpus comprising lots of equal-sized text segments (e.g. the LOB and Brown corpora). But in the real world, especially if your research focus is the text as opposed to the language, you will probably be dealing with texts of different lengths and the conventional TTR will not help you much.

WordList offers a better strategy as well: the *standardised type/token ratio* (STTR) is computed every **n** words as Wordlist goes through each text file. By default 113, n = 1,000. In other words the ratio is calculated for the first 1,000 running words, then calculated afresh for the next 1,000, and so on to the end of your text or corpus. A running average is computed, which means that you get an average type/token ratio based on consecutive 1,000-word chunks of text. (Texts with less than 1,000 words (or whatever n is set to) will get a standardised type/token ratio of 0.)

Setting the N boundary

Adjust the n number in Minimum & Maximum Settings 314 to any number between 100 and 20,000.

What STTR actually counts

Note: The ratio is computed a) counting every different form 270 as a word (so **say** and **says** are two types) b) using only the words which are not in a stop-list 120 c) those which are within the length you have specified, d) taking your preferences about numbers 446 and hyphens 435 into account.

The number shown is a percentage of new types for every n tokens. That way you can compare type/token ratios across texts of differing lengths. This method contrasts with that of Tuldava 417 (1995:131-50) who relies on a notion of 3 stages of accumulation. The WordSmith method of computing STTR was my own invention but parallels one of the methods devised by the mathematician David Malvern working with Brian Richards (University of Reading).

Further discussion

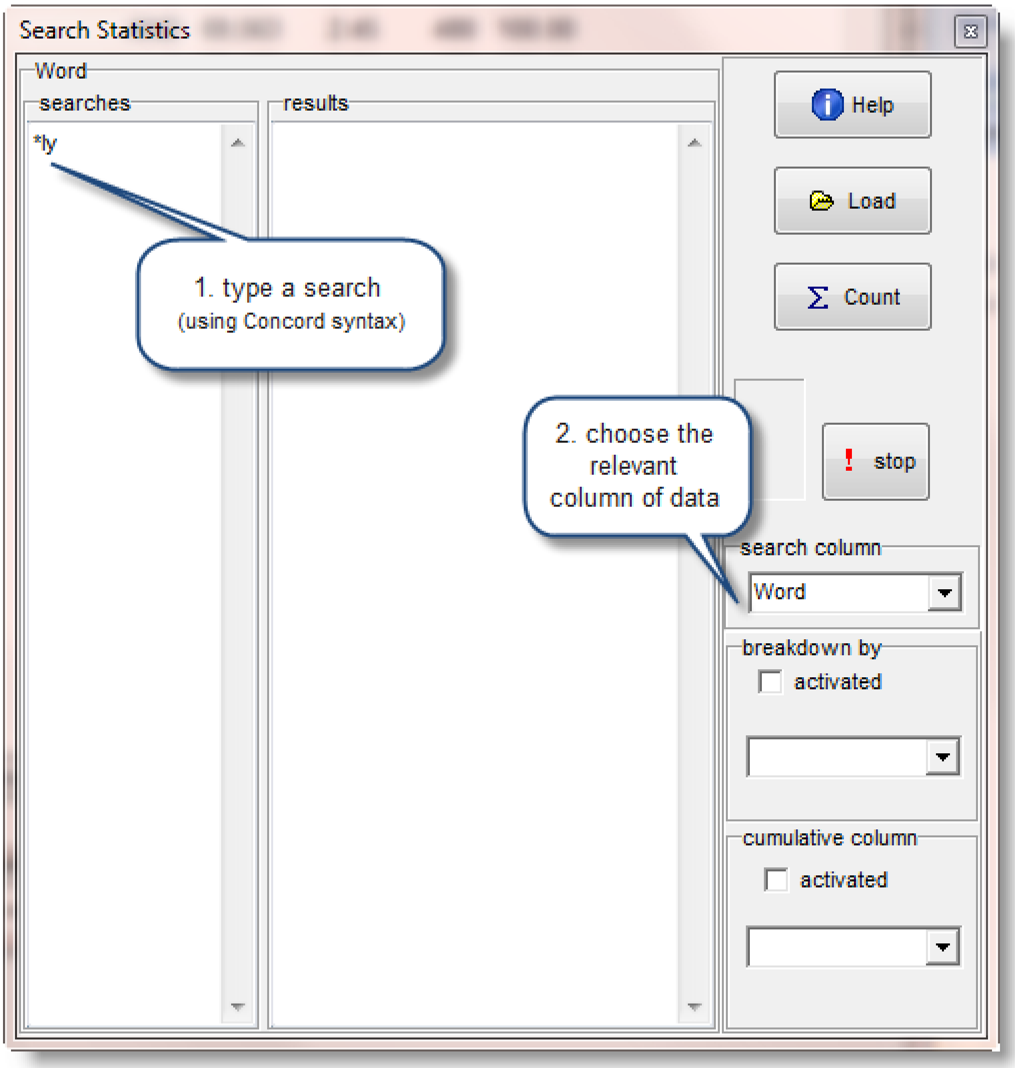
TTR and STTR are both pretty crude measures even if they are often assumed to imply something about "lexical density". Suppose you had a text which spent 1,000 words discussing **ELEPHANT, LION, TIGER** etc, and then 1,000 discussing **MADONNA, ELVIS**, etc., then 1,000 discussing **CLOUD, RAIN, SUNSHINE**. If you set the STTR boundary at 1,000 and happened to get say 48% or so for each section, the statistic in itself would not tell you there was a change involving Africa, Music, Weather. Suppose the boundary between Africa & Music came at word 650 instead of at word 1,000, I guess there'd be little or no difference in the statistic. But what *would* make a difference? A text which discussed clouds and written by a person who distinguished a lot between types of cloud might also use **MIST, FOG, CUMULUS, CUMULO-NIMBUS.** This would be higher in STTR than one written by a child who kept referring to **CLOUD** but used adjectives like **HIGH, LOW, HEAVY, DARK, THIN, VERY THIN** to describe the clouds... and who repeated **DARK, THIN**, etc a lot in describing them.....

(NB. Shakespeare is well known to have used a rather limited vocabulary in terms of measures like these!)

9.12.3 summary statistics

A word list's statistics give you data about the corpus, but you may need more specific information about individual words in a word list too.

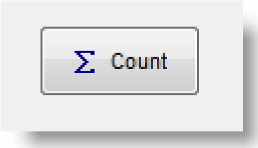
How many end in -ly?



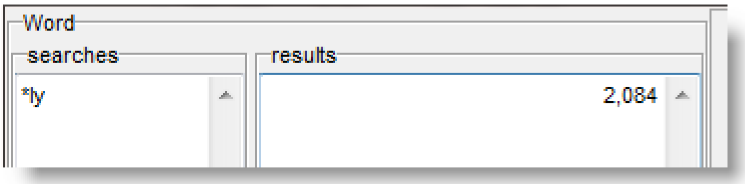
1. type a search (using Concord syntax)

2. choose the relevant column of data

Press Count

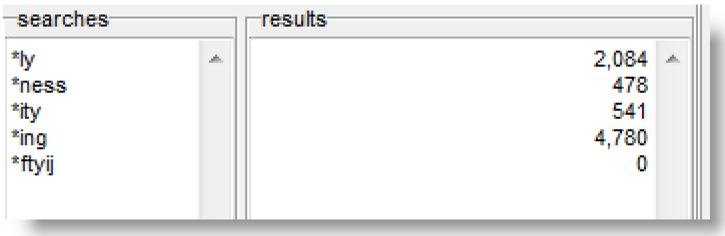


to get something like this:



searches	results
*ly	2,084

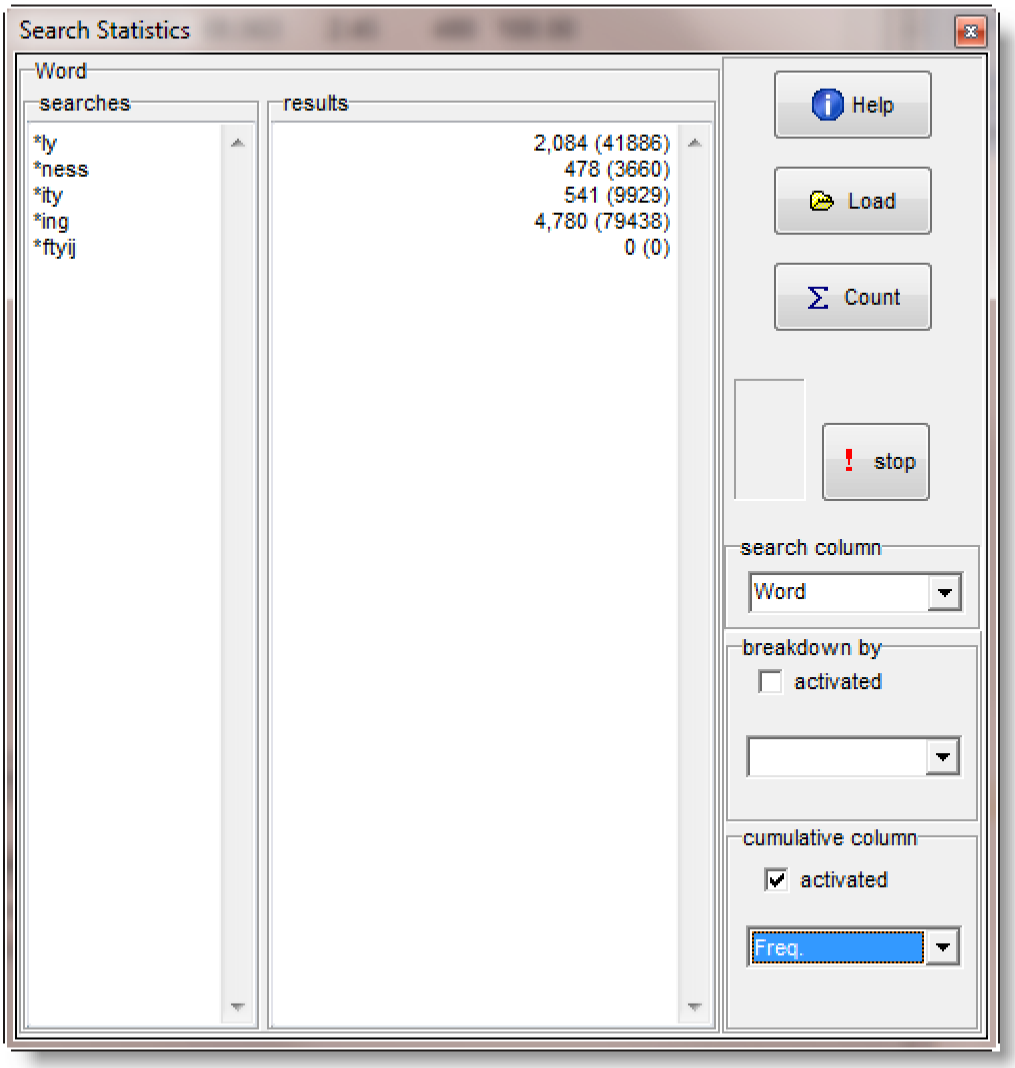
There is no limit on the searches:



searches	results
*ly	2,084
*ness	478
*ity	541
*ing	4,780
*ftyij	0

Cumulative Column

A cumulative count adds up scores on another column of data apart from the one you are processing for your search. The columns in this window are for numerical data only. Select one and ensure *activated* is ticked.



Search Statistics

Word

searches	results
*ly	2,084 (41886)
*ness	478 (3660)
*ity	541 (9929)
*ing	4,780 (79438)
*ftyij	0 (0)

search column: Word

breakdown by: □ activated

cumulative column: ☑ activated Freq.

In this example, a word-list was computed and a search was made of 4 common word endings (and one ridiculous one). For **-LY** there are 2,084 types, with a total of 41,886 tokens in this corpus. **-ITY** and **-NESS** are found at the ends of fairly similar numbers of word-types, but **-ITY** has many more tokens in these data.

Breakdown

See the example for Concord 215

Load button

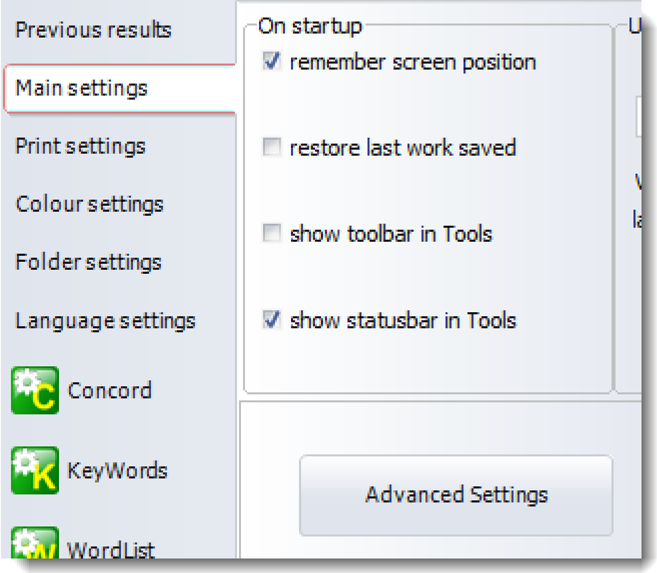
see the explanation for count data frequencies 66.

9.13 stop-lists and match-lists

In WordList, a stop list 120 is used in order to filter out some words, usually high-frequency words, that you want excluded from your word-list. The idea of a match-list 92 is to be able to compare all the words in your word list with another list in a plain text file and then do one of a variety of operations such as deleting the words which match, deleting those which don't, or just marking the ones in the list.

For both, you can define your own lists and save them in plain text files.

Settings are accessed via the WordList menu or by an Advanced Settings button in the Controller



On startup

☑ remember screen position

☐ restore last work saved

☐ show toolbar in Tools

☑ show statusbar in Tools

See also: lemma lists 270, general explanation of stop-lists 120

9.14 import words from text list

the point of it

You might want a word list based on some data you have obtained in the form of a list, but whose original texts you do not have access to.

requirements

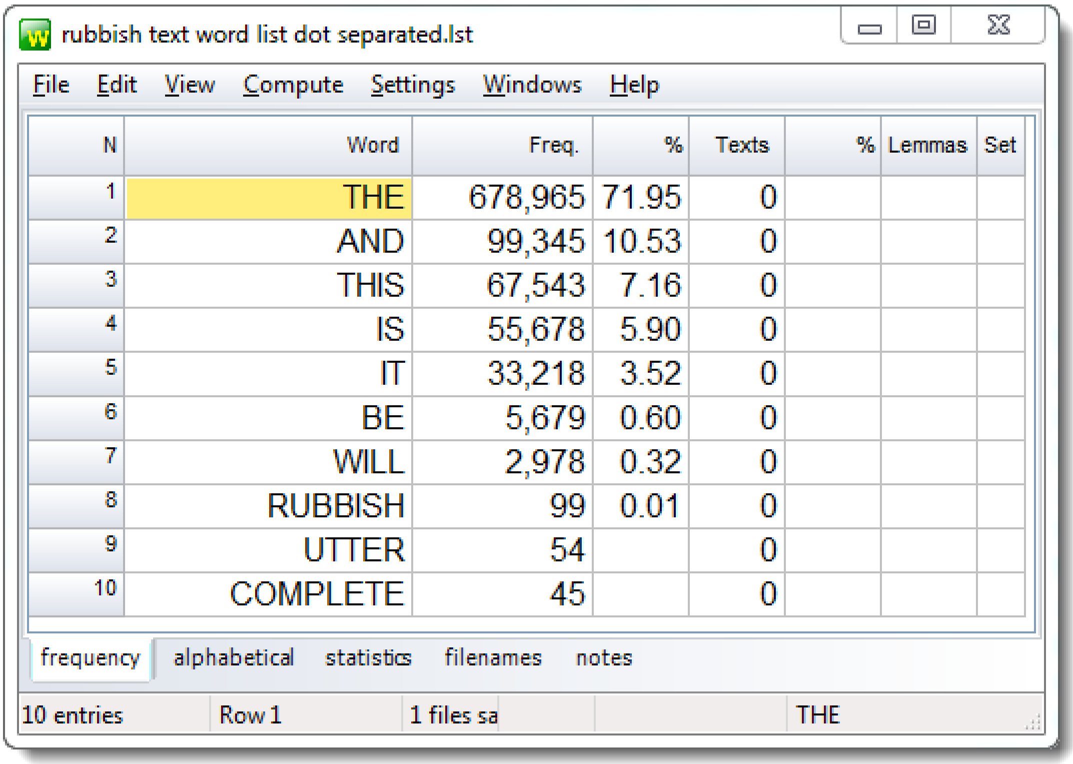
Your text file can be in any language 81 (select this before you make the list), and can be in Unicode or ASCII or ANSI, plain text.
<Tab> characters are expected to separate the columns of data. Decimal points and commas will be ignored. Words will have leading or trailing spaces trimmed off. The words do not need to be in frequency or alphabetical order. You need at least a column with words and another with a number representing each word's frequency.

example

```
; My word list for test purposes.

THIS     67,543
```

```
IT        33,218
WILL      2,978
BE        5,679
COMPLETE  45
AND       99,345
UTTER     54
RUBBISH   99
IS        55,678
THE       678,965
```

You should get results like these.



N	Word	Freq.	%	Texts	%	Lemmas	Set
1	THE	678,965	71.95	0			
2	AND	99,345	10.53	0			
3	THIS	67,543	7.16	0			
4	IS	55,678	5.90	0			
5	IT	33,218	3.52	0			
6	BE	5,679	0.60	0			
7	WILL	2,978	0.32	0			
8	RUBBISH	99	0.01	0			
9	UTTER	54		0			
10	COMPLETE	45		0			

10 entries | Row 1 | 1 files sa | THE

Statistics are calculated in the simplest possible way: the word-lengths (plus mean and standard deviation), and the number of types and tokens. Most procedures need to know the total number of running words (tokens) and the number of different word types so you should manage to use the word-list in KeyWords etc.

The total is computed by adding the frequencies of each word-type (`67543+33218+2978` etc. in the example above).

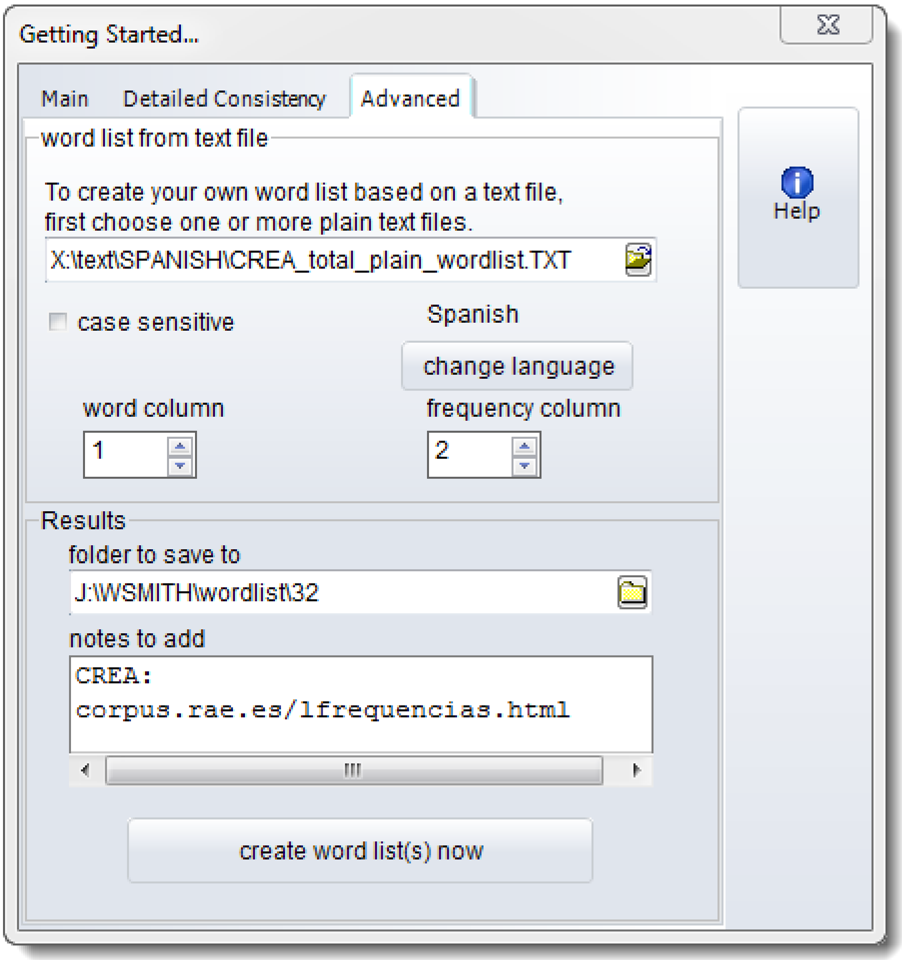
Optionally, a line can start **`\TOTAL=\`** and contain a numerical total, eg.

```
\TOTAL=\  299981
```

In this case the total number of tokens will be assumed to be 299981, instead.

how to do it

When you choose the *New* menu option (●) in WordList you get a window offering three tabs: a *Main* tab for most usual purposes,



To create your own word list based on a text file,
first choose one or more plain text files.

X:\text\SPANISH\CREA_total_plain_wordlist.TXT

□ case sensitive

Spanish

word column: 1

frequency column: 2

folder to save to: J:\WSMITH\wordlist\32

notes to add:

```
CREA:
corpus.rae.es/lfrequencias.html
```

one for *Detailed Consistency* 263, and another (*Advanced*) for creating a word list using a plain text file.

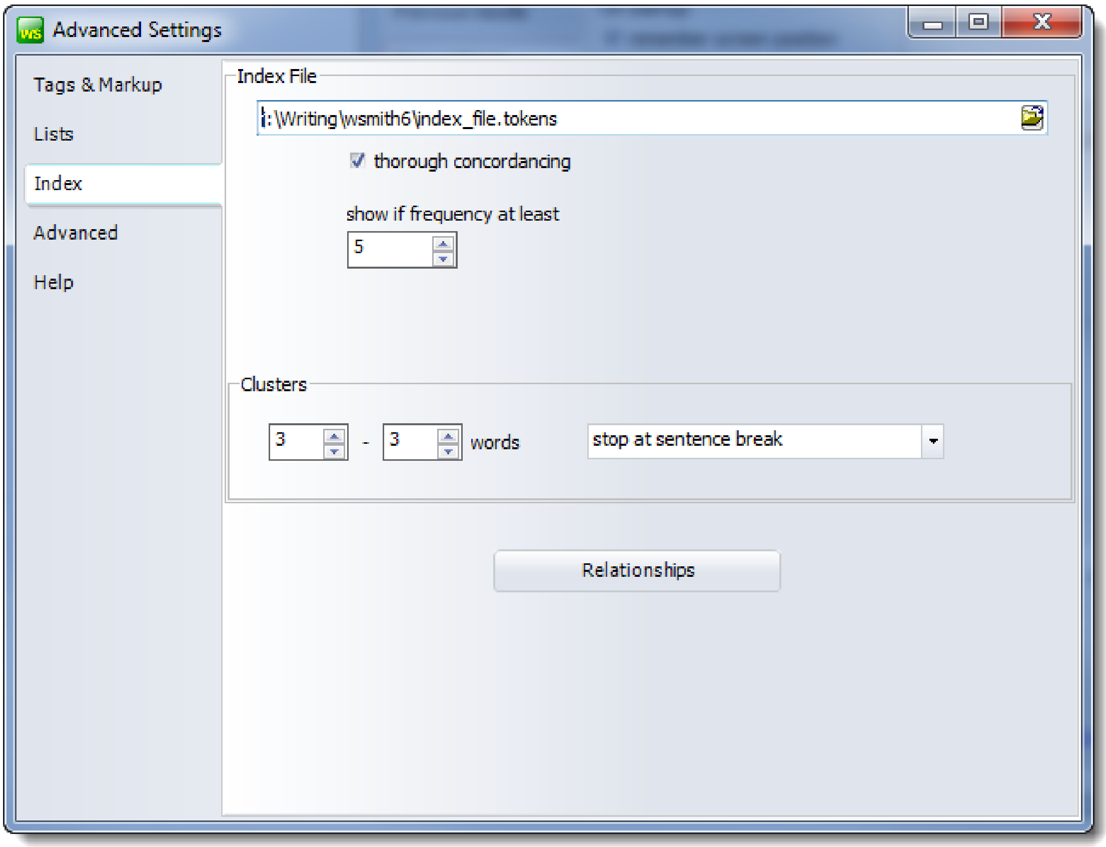
Set the *word column* and *frequency column* appropriately according to the tabs in each line. (Column 1 assumes that the word comes first before any tabs; in the case of CREA's Spanish word-list there is a column for ranking so the word and frequency columns would need to be 2 and 3 respectively.)

Choose your .txt file(s) and a suitable folder to save to, add any notes you wish, and press *create word list(s) now.*

9.15 settings

Enter topic text here.

9.15.1 WordSmith controller: Index settings



Index File

i:\Writing\wsmith6\index_file.tokens

☑ thorough concordancing

show if frequency at least: 5

Clusters: 3 - 3 words

stop at sentence break

Index File

The filename is for a default index which you wish to consider the usual one to open.

thorough concordancing: when you compute a concordance from an index, you will either get (*thorough* checked) or not get (if not checked) full sentence, paragraph and other statistics 166 as in a normal concordance search. (Computing these statistics takes a little longer.)

show if frequency at least: determines which items you will see when you load up the index file. (What you see looks like a word list but it is reading the underlying index.)

Clusters

the minimum and maximum sizes are 2 and 8. Set these before you compute a multi-word word list 278 based on the index. A good maximum is probably 5 or 6.

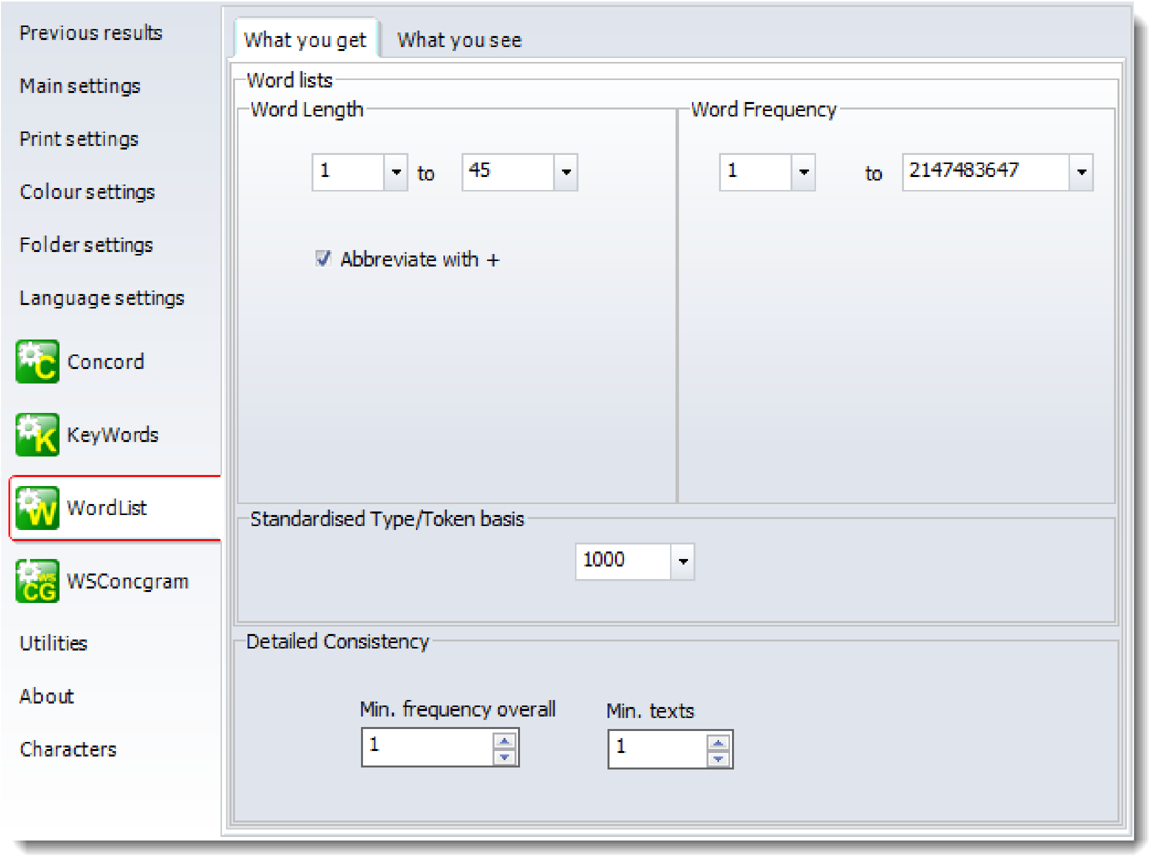
stop at: you can choose where you want cluster breaks to be assumed. With the setting above (no limits), "`I wrote the letter. Then I posted it`" would consider `letter then I posted` as a possible multi-word string even though there's a sentence break 188 between them.

Relationships

See relationships computing 294.

9.15.2 WordSmith controller: WordList settings

These are found in the main Controller 4 marked *WordList* .
This is because some of the choices -- e.g. Minimum & Maximum Settings 314 -- may affect other Tools.
There are 2 sets : *What you Get* and *What you See*.



Word lists

Word Length: 1 to 45

☑ Abbreviate with +

Word Frequency: 1 to 2147483647

Standardised Type/Token basis: 1000

Detailed Consistency

Min. frequency overall: 1

Min. texts: 1

WHAT YOU GET

Word Length & Frequencies

See Minimum & Maximum Settings 314.

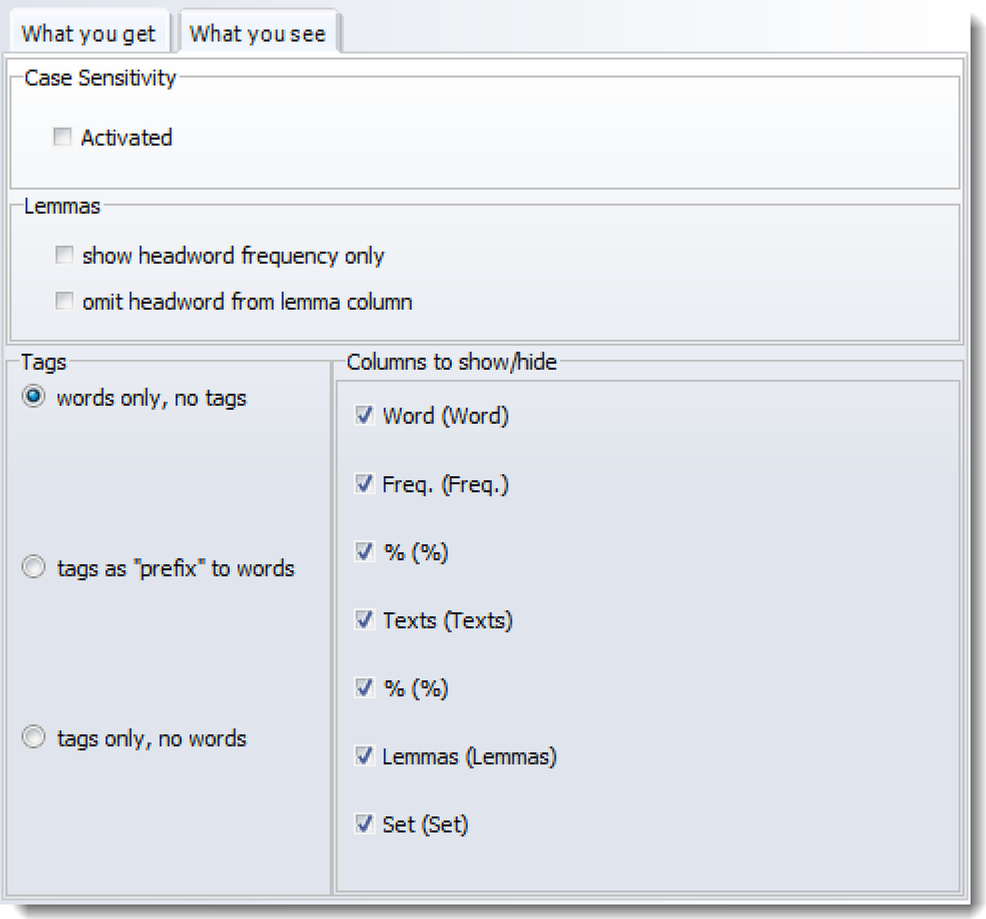
Standardised Type/Token

See WordList Type/Token Information 303.

Detailed Consistency

Min. frequency overall = a total frequency for a word to be included in the Detailed Consistency 263 list.
Min. texts = the minimum number of texts that word must appear in.

WHAT YOU SEE



Case Sensitivity

□ Activated

Lemmas

□ show headword frequency only

□ omit headword from lemma column

Tags

● words only, no tags

○ tags as "prefix" to words

○ tags only, no words

Columns to show/hide

☑ Word (Word)

☑ Freq. (Freq.)

☑ % (%)

☑ Texts (Texts)

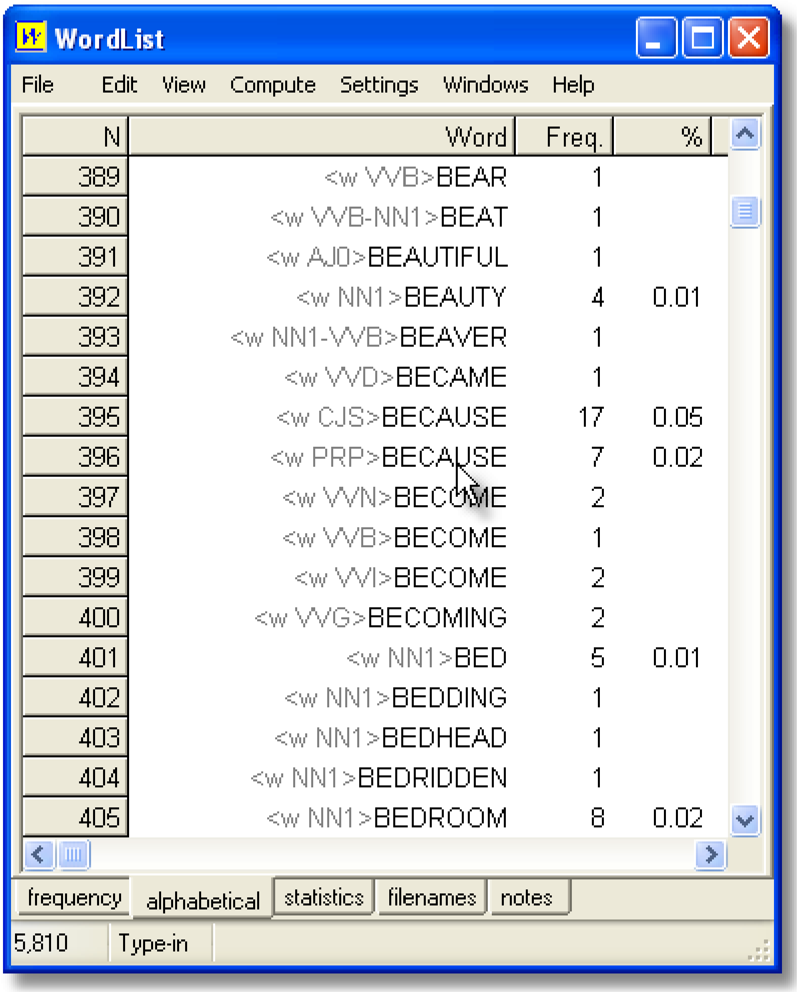
☑ % (%)

☑ Lemmas (Lemmas)

☑ Set (Set)

Tags

By default you get "words only, no tags". If you want to include tags in a word list, you need to set up a Tag File 141 first. Then choose one of the options here.
In the example here we see that **BECAUSE** is classified by the BNC either as a **<w CJS>** or a **<w PRP>**. (That's how the BNC classifies **BECAUSE OF**...)



N	Word	Freq.	%
389	<w VVB>BEAR	1	
390	<w VVB-NN1>BEAT	1	
391	<w AJ0>BEAUTIFUL	1	
392	<w NN1>BEAUTY	4	0.01
393	<w NN1-VVB>BEAVER	1	
394	<w VVD>BECAME	1	
395	<w CJS>BECAUSE	17	0.05
396	<w PRP>BECAUSE	7	0.02
397	<w VVN>BECOME	2	
398	<w VVB>BECOME	1	
399	<w VVI>BECOME	2	
400	<w VVG>BECOMING	2	
401	<w NN1>BED	5	0.01
402	<w NN1>BEDDING	1	
403	<w NN1>BEDHEAD	1	
404	<w NN1>BEDRIDDEN	1	
405	<w NN1>BEDROOM	8	0.02

For colours and tags see WordList and Tags 315.

Columns

The *Columns to show/hide* list offers all the standard columns: you may uncheck ones you normally do not wish to see. This will only affect newly computed KeyWords data: earlier data uses the column visibility, size, colours etc already saved. They can be altered using the Layout 87 menu option at any time.

Case Sensitivity

Normally, you'll make a case-insensitive word list. If you wish to make a word list which distinguishes between `the`, `The` and `THE`, activate case sensitivity 314.

Lemma Visibility

By default in a word-list you'll see the frequency of the headword plus the associated forms; if you check the *show headword frequency only* box, the frequency column will ignore the associated wordform frequencies. Similarly, if you check *omit headword from lemma column* you will see only the associated forms there.

See also: Using Index Lists 276, Viewing Index Lists 284, WordList Help Contents 258, WordList and tags 315, Computing word list clusters 278.

9.15.3 minimum & maximum settings

These include:

minimum word length

Default: 1 letter. When making a word-list, you can specify a minimum word length, e.g. so as to cut out all words of less than 3 letters.

maximum word length

Default: 49 letters. You can allow for words of up to 50 characters in length. If a word exceeds the limit and Abbreviate with + is checked, WordList will append a + symbol at the end of it to show that it was cut short. (If Abbreviate with + is not checked, the long word will be omitted from your word list. You might wish to use this to set both minimum and maximum to say, 4, and leave Abbreviate with + un-checked – that way you'll get a word-list with only the 4-letter words in it.

minimum frequency

Default: 1. By default, all words will be stored, even those which occur once only. If you want only the more frequent words, set this to any number up to 32,000.

maximum frequency

Default maximum is 2,147,483,647 (2 Gigabytes). You'd have to analyse a lot of text to get a word which occurred as frequently as that!. You might set this to say 500, and the minimum to 50: that way your word-list would hold only the moderately common words.

type/token mean number (default 1,000)

Enables a smoothed calculation of type/token ratio for word lists. Choose a number between 10 and 20,000. For a more complete explanation, see WordList Type/Token Information 303.

See also: Text Characteristics 124, Stop Lists 120, Setting Defaults 113

9.15.4 case sensitivity

Normally, you'll make a case-insensitive word list, especially as in most languages capital letters are used not only to distinguish proper nouns but also to signal beginnings of sentences, headings, etc. If, however, you wish to make a word list which distinguishes between major, Major and MAJOR, activate case sensitivity (*WordList Settings* | *Case Sensitivity* in the Controller 4).

When you first see your case-sensitive list, it is likely to appear all in UPPER CASE. Press Ctrl+L or choose the *Layout* 87 menu option (▥) to change this.

9.16 sorting

How to do it...

Sorting can be done simply by pressing the top row of any list. Press again to toggle between ascending & descending sorts.

With a word-list on your screen, the main Frequency window doesn't sort, but you can re-sort the Alphabetical window (look at the tabs at the bottom of WordList to choose the tab) in a number of different ways.

The menu offers various options.

Alphabetical Word Sort

Many languages have their own special sorting order, so prior to sorting or re-sorting, check that you have selected the right language 81 for the words being sorted. Spanish, for example, uses this order: A,B,C,CH,D,E,F,G,H,I,J,K,L,LL,M,N,Ñ,O,P,Q,R,S,T,U,V,W,X,Y,Z.

KeyWords and other comparisons require an alphabetically-ordered list in ascending order. If you get problems, please open the word lists in WordList, choose the "alphabetical" tab, sort by pressing the "Word" header until the sort is definitely alphabetical ascending, then choose the Save menu option.

Reverse Word Sort (Ctrl+F6)

This is so that you can sort words by suffix. The order is determined by word endings, not word beginnings. You will therefore find all the `-ing` forms together.

Word Length Sort (Shift+Ctrl+F6)

This is so that you can sort words by their length (1-letter, 2-letter, etc up to 50-letter words) Within a set of equal-length words, there's a second, alphabetical sort.

Consistency Sort

Press the "Texts" header to re-sort the words according to their consistency 263.

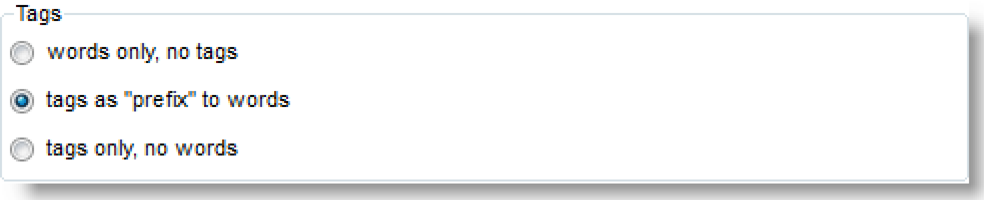
See also: Concord sort 208, KeyWords sort 252, Editing entries 72; Accented characters 419; Choosing Language 81

9.17 WordList and tags

If you have defined a tag file and made the appropriate settings 311 to load it, you can get a word-list which treats tags and words separately as in this example, where the tag is viewed as if it were a prefix.

A word list only of tags?

Choose whether you want only the tags, only the words or both in *WordList settings* | *What you see* | *Tags*:



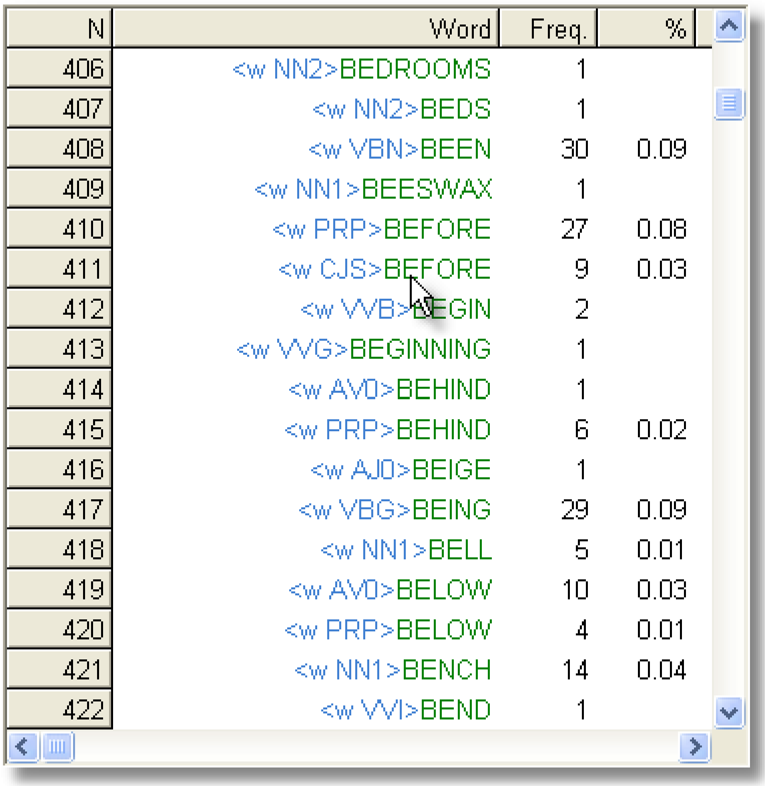
Tags

○ words only, no tags

● tags as "prefix" to words

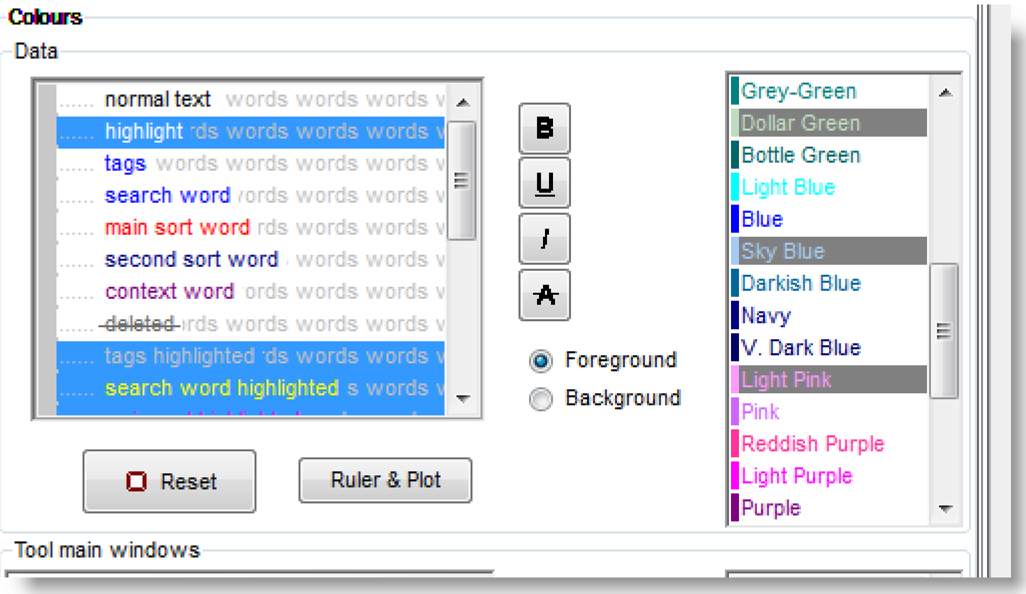
○ tags only, no words

In its *Alphabetical* view, the list can be sorted on the tag or the word.

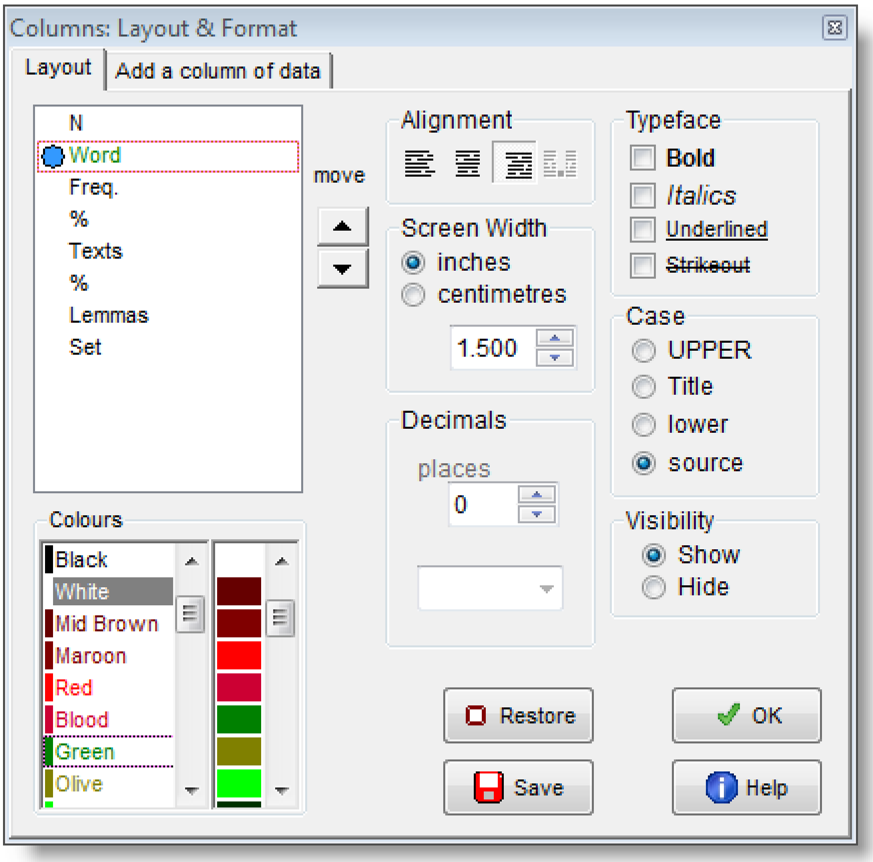


N	Word	Freq.	%
406	<w NN2>BEDROOMS	1	
407	<w NN2>BEDS	1	
408	<w VBN>BEEN	30	0.09
409	<w NN1>BEESWAX	1	
410	<w PRP>BEFORE	27	0.08
411	<w CJS>BEFORE	9	0.03
412	<w VVB>BEGIN	2	
413	<w VVG>BEGINNING	1	
414	<w AV0>BEHIND	1	
415	<w PRP>BEHIND	6	0.02
416	<w AJ0>BEIGE	1	
417	<w VBG>BEING	29	0.09
418	<w NN1>BELL	5	0.01
419	<w AV0>BELOW	10	0.03
420	<w PRP>BELOW	4	0.01
421	<w NN1>BENCH	14	0.04
422	<w VVI>BEND	1	

To colour these as in the example, in the main Controller I chose Blue for the foreground for tags (as the default is a light grey).



Then in WordList, I chose *View* | *Layout* as in this screenshot, selected the Word column header and chose green below.

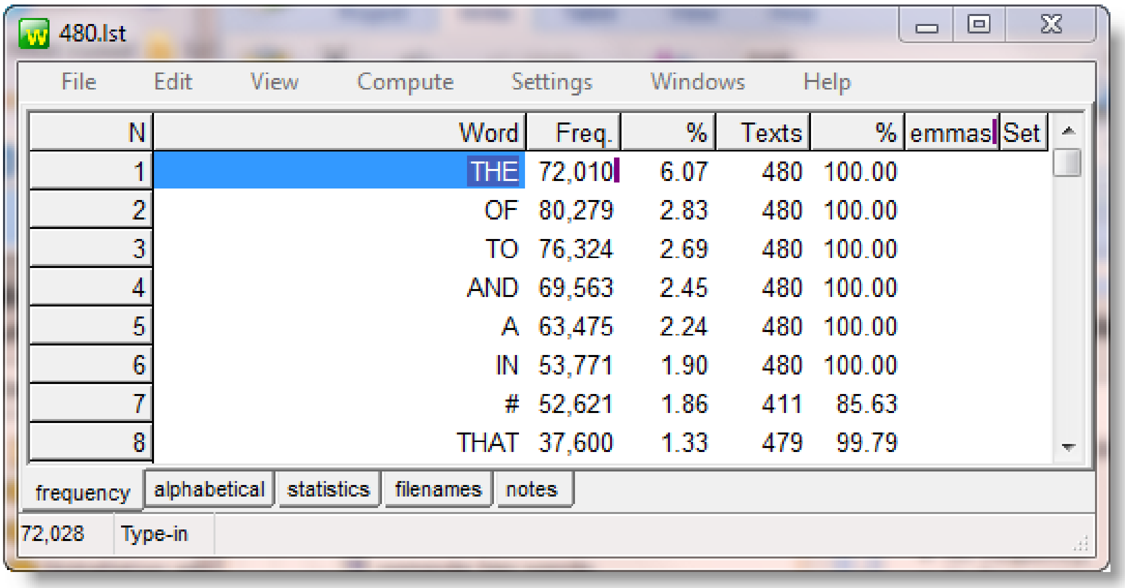


9.18 WordList display

Each WordList display shows

- the word
- its frequency
- its frequency as a percent of the running words in the text(s) the word list was made from
- the number of texts each word appeared in
- that number as a percentage of the whole corpus of texts

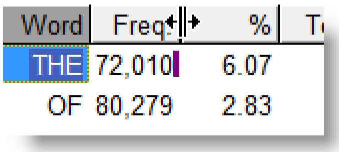
The Frequency display might look like this:



N	Word	Freq.	%	Texts	%	emmas	Set
1	THE	72,010	6.07	480	100.00		
2	OF	80,279	2.83	480	100.00		
3	TO	76,324	2.69	480	100.00		
4	AND	69,563	2.45	480	100.00		
5	A	63,475	2.24	480	100.00		
6	IN	53,771	1.90	480	100.00		
7	#	52,621	1.86	411	85.63		
8	THAT	37,600	1.33	479	99.79		

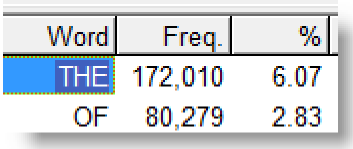
Here you see the top 7 word-types in a word list based on 480 texts. There are 72,028 occurrences of these words (tokens 426) altogether but in the screenshot we can only see the first few. The Freq. column shows how often each word cropped up (**THE** *looks* as if it appeared 72,010 times in the 480 texts), and the % column tells us that the frequency represents 6.07% of the running words in those texts. The Texts column shows that **THE** comes in 480 texts, that is 100% of the texts used for the word list.

If we pull the *Freq.* column a little wider



Word	Freq.	%	T
THE	72,010	6.07	
OF	80,279	2.83	

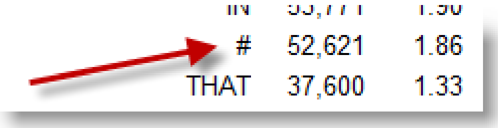
(cursor at the header edge then pull right) so that the 72,010 doesn't have any purple marks beside it,



Word	Freq.	%
THE	172,010	6.07
OF	80,279	2.83

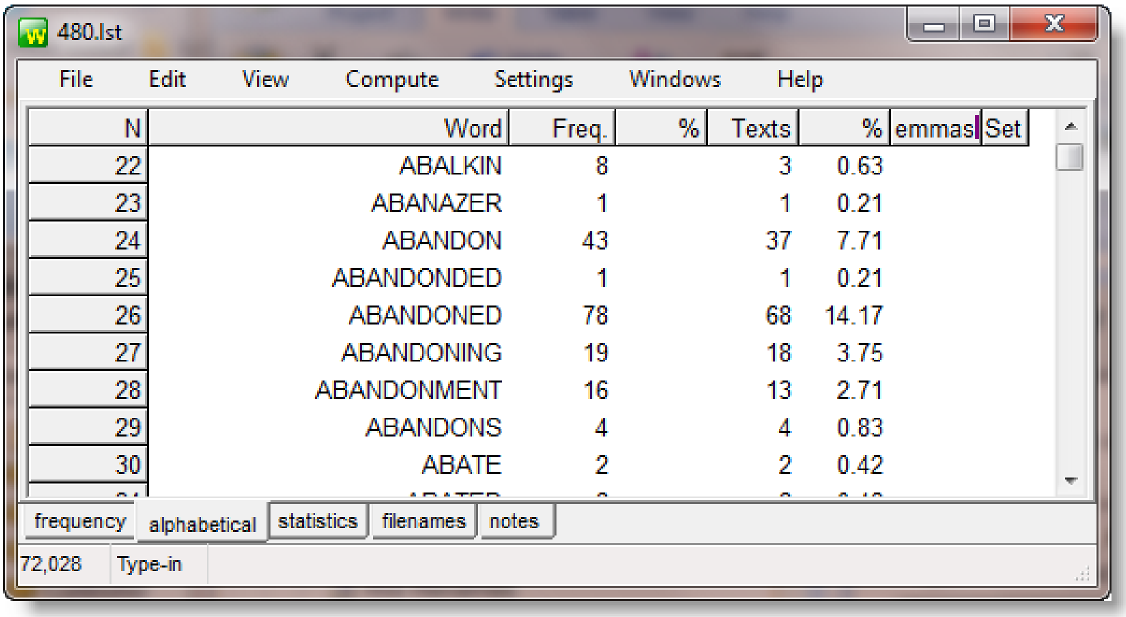
we see the true frequency value is actually 172,010.

Another thing to note is that there seems to be a word #, with over 50 thousand occurrences.



IN	53,771	1.90
#	52,621	1.86
THAT	37,600	1.33

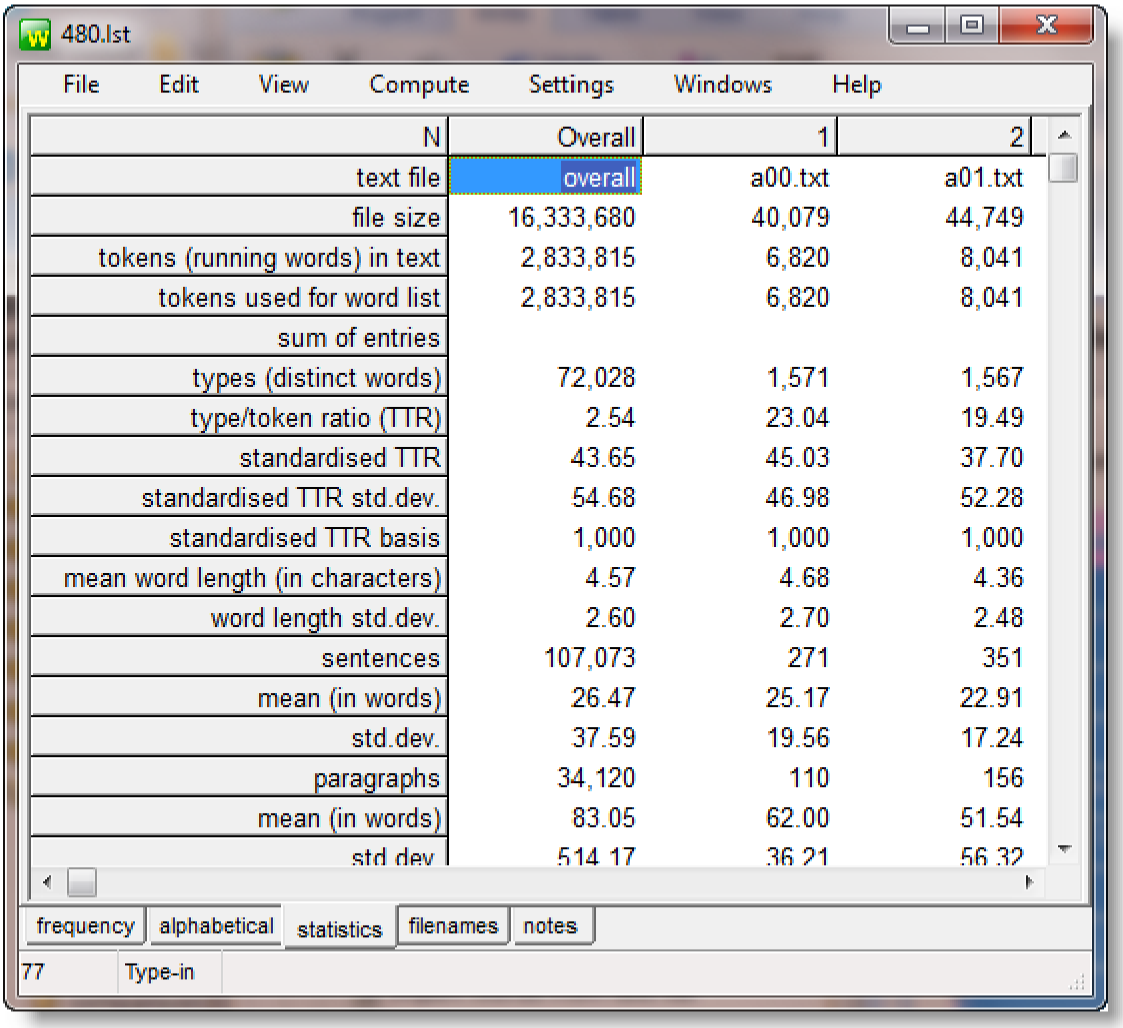
That represents a number or any word with a number in it such as **EX658**.



N	Word	Freq.	%	Texts	%	emmas	Set
22	ABALKIN	8		3	0.63		
23	ABANAZER	1		1	0.21		
24	ABANDON	43		37	7.71		
25	ABANDONDED	1		1	0.21		
26	ABANDONED	78		68	14.17		
27	ABANDONING	19		18	3.75		
28	ABANDONMENT	16		13	2.71		
29	ABANDONS	4		4	0.83		
30	ABATE	2		2	0.42		

The Alphabetical listing also shows us some of the words but now they're in alphabetical order. **ABANDON** comes 43 times altogether, and in 37 of the 480 texts (less than 8%). **ABANDONED**, on the other hand, not only comes more often (78 times) but also in more texts (14% of them).

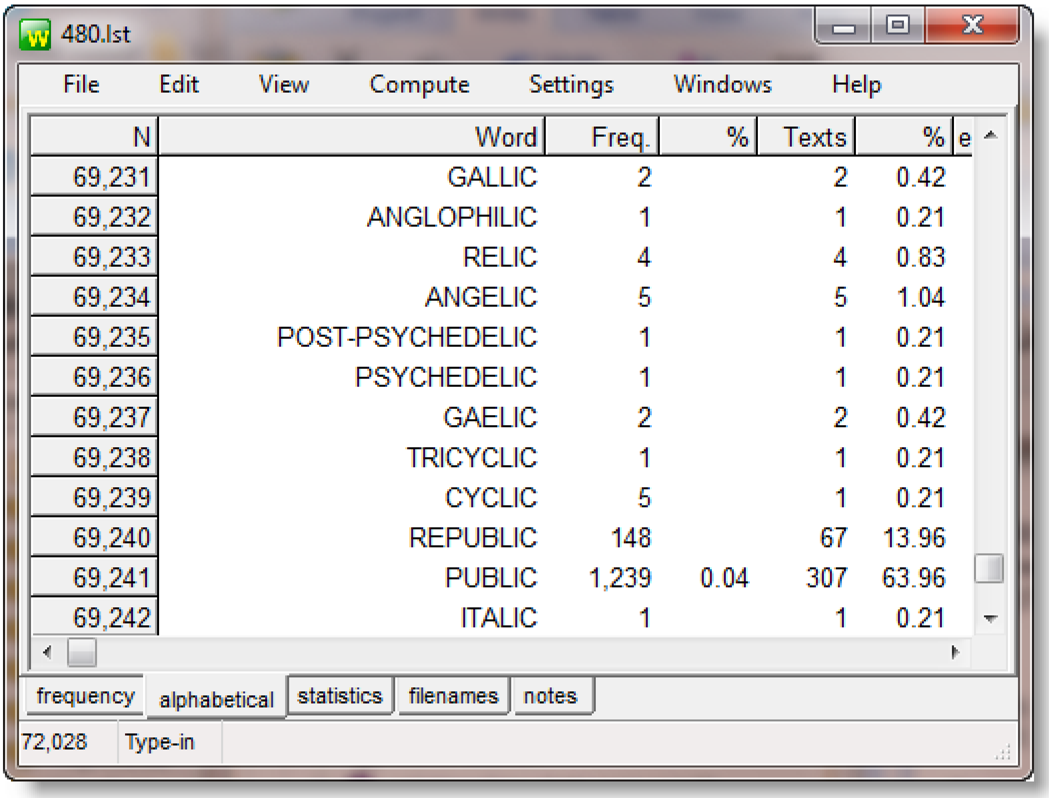
Now let's examine the statistics.



N	Overall	1	2
text file	overall	a00.txt	a01.txt
file size	16,333,680	40,079	44,749
tokens (running words) in text	2,833,815	6,820	8,041
tokens used for word list	2,833,815	6,820	8,041
sum of entries			
types (distinct words)	72,028	1,571	1,567
type/token ratio (TTR)	2.54	23.04	19.49
standardised TTR	43.65	45.03	37.70
standardised TTR std.dev.	54.68	46.98	52.28
standardised TTR basis	1,000	1,000	1,000
mean word length (in characters)	4.57	4.68	4.36
word length std.dev.	2.60	2.70	2.48
sentences	107,073	271	351
mean (in words)	26.47	25.17	22.91
std.dev.	37.59	19.56	17.24
paragraphs	34,120	110	156
mean (in words)	83.05	62.00	51.54
std.dev.	514.17	36.21	56.32

In all 480 texts, there are 72,028 word types (as pointed out above). The total running words is 2,833,815. Each word is about 4.57 characters in length. There are 107,073 sentences altogether, on average 26.47 words in length. In the text of **a00.txt**, there are only 1,571 different word types and that interview is under 7,000 words in length. This is explained in more detail in the Statistics 298 page.

Finally, here is a screenshot of the same word list sorted "reverse alphabetically". In the part which we can see, all the words end in **-IC**.



N	Word	Freq.	%	Texts	%
69,231	GALLIC	2		2	0.42
69,232	ANGLOPHILIC	1		1	0.21
69,233	RELIC	4		4	0.83
69,234	ANGELIC	5		5	1.04
69,235	POST-PSYCHEDELIC	1		1	0.21
69,236	PSYCHEDELIC	1		1	0.21
69,237	GAELIC	2		2	0.42
69,238	TRICYCLIC	1		1	0.21
69,239	CYCLIC	5		1	0.21
69,240	REPUBLIC	148		67	13.96
69,241	PUBLIC	1,239	0.04	307	63.96
69,242	ITALIC	1		1	0.21

To do a reverse alphabetical sort, I had the Alphabetical window visible, then chose *Edit* | *Other sorts* | *Reverse Word sort* in the menu. To revert to an ordinary alphabetical sort, press F6.

See also : Consistency 262, Lemmatisation 270

Utility Programs

Section X

10 Utility Programs

Besides the three main programs, there are more Tools that have arisen over the years; this Chapter explains them.

Character Profiler 405	lists characters used in your texts
CharGrams 7	like WordList but for sequences of characters
Corpus Corruption Detector 7	to find anomalous texts
File Utilities 8	various utilities for managing files
File Viewer 8	shows the innards of your text files
Minimal Pairs 8	identifies similar words
Text Converter 354	prepares your corpora for different formats
Viewer and Aligner 379	shows translated texts
WSConcGram 12	finds and shows concgrams

10.1 Convert Data from Previous Versions

10.1.1 Convert Data from Previous Versions

As WordSmith Tools develops, it has become necessary to store more data along with any given word-list, concordance etc. For example, data about which language 81 (s) were selected for a concordance, notes 29 now stored with every type of results file, etc. Therefore it has been necessary to supply a tool to convert data from the formats used in WS 1.0 to 3.0 (last millennium) to the new format for the current version.

This is the Data Converting tool.

If you try to open a file made with a previous version you should be offered a chance to convert it first.

Note: as WordSmith develops, its saved data may get more complex in format. A concordance saved by WordSmith 5.0 cannot be guaranteed to be readable by WordSmith 4.0 for that reason, and a 6.0 one may require version 6.0, etc.

10.2 WebGetter

10.2.1 overview

The point of it

The idea is to build up your own corpus of texts, by downloading web pages with the help of a search engine.

What you do

Just type a word or phrase, check the language, and press *Download*.

How it works

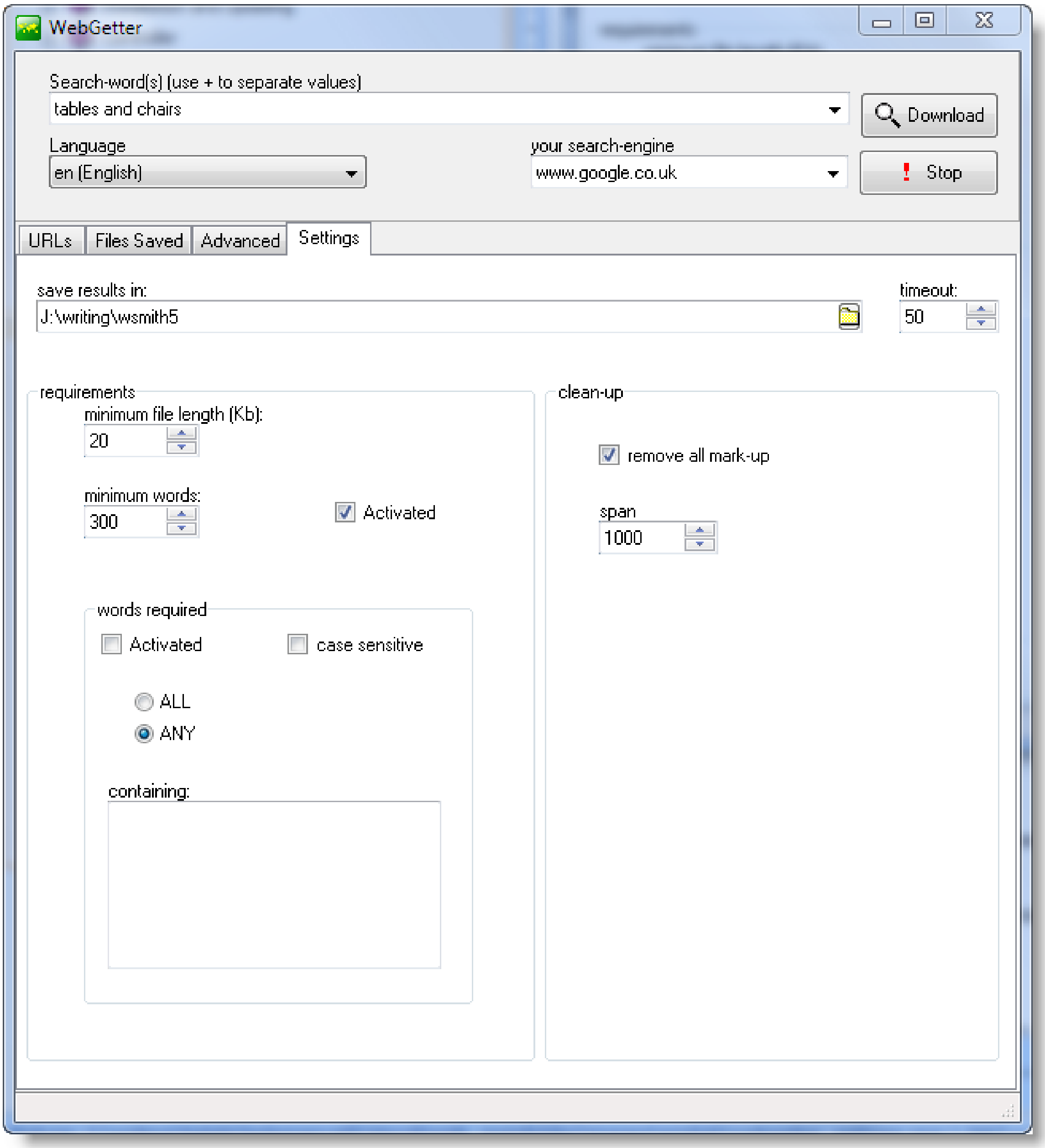
WebGetter visits the search engine you specify and downloads the first 1000 sources or so. Basically it uses the search engine just as you do yourself, getting a list of useful references. Then it sends out a robot to visit each web address and download the web page in each case (not from the search engine's cache but from the original web-site). Quite a few robots may be out there searching for you at once -- the advantage of this is that one slow download doesn't hold all the others up.

After downloading a web page, that WebGetter robot checks it meets your requirements (in Settings 325) and cleans up the resulting text. If the page is big enough, a file with a name very similar to the web address will be saved to your hard disk.

When it runs out of references, **WebGetter** re-visits the search engine and gets some more.

See also: Settings 325, Display 326, Limitations 328

10.2.2 settings



WebGetter

Search-word(s) (use + to separate values): tables and chairs

Language: en (English)

your search-engine: www.google.co.uk

save results in: J:\writing\wsmith5

timeout: 50

requirements

minimum file length (Kb): 20

minimum words: 300

☑ Activated

words required

□ Activated □ case sensitive

○ ALL

● ANY

containing:

clean-up

☑ remove all mark-up

span: 1000

Language

Choose the language you require from the drop-down list.

Search Engine

The search engine box allows you to choose for example www.google.com.br for searches on Brazilian Portuguese or www.google.fr for French. That is a better guarantee of getting text in the language you require!

Folder and Time-out

- where the texts are to be stored. By defaults it is the **`\wsmith5`** folder stemming from your **`My Documents`**. The folder you specify will act as a root. That is, if you specify **`c:\temp`** and search for "besteirol", results will be stored in **`c:\temp\besteirol`**. If you do another search on say "WordSmith Tools", results for that will go into **`c:\temp\WordSmith Tools`**.
- timeout: the number of seconds after which **WebGetter** robot stops trying a given webpage if there's no response. Suggested value: 50 seconds.

Requirements

- minimum file length (suggested 20Kbytes): the minimum size for each text file downloaded from the web. Small ones may just contain links to a couple of pictures and nothing much else.
- minimum words (suggested: 300): after each download, **WebGetter** goes through the downloaded text file counting the number of words and won't save unless there are enough.
- required words: you may optionally type in some words which you require to be present in each download; you can insist they all be present or any 1 of these.

Clean-up

If you want all the HTML markup removed, you can check this box, setting a suitable span between < and > markers, 1000 recommended.

Advanced Options

If you work in an environment with a "Proxy Server", **WebGetter** will recognise this automatically and use the proxy unless you uncheck the relevant box. If in doubt ask your network administrator.

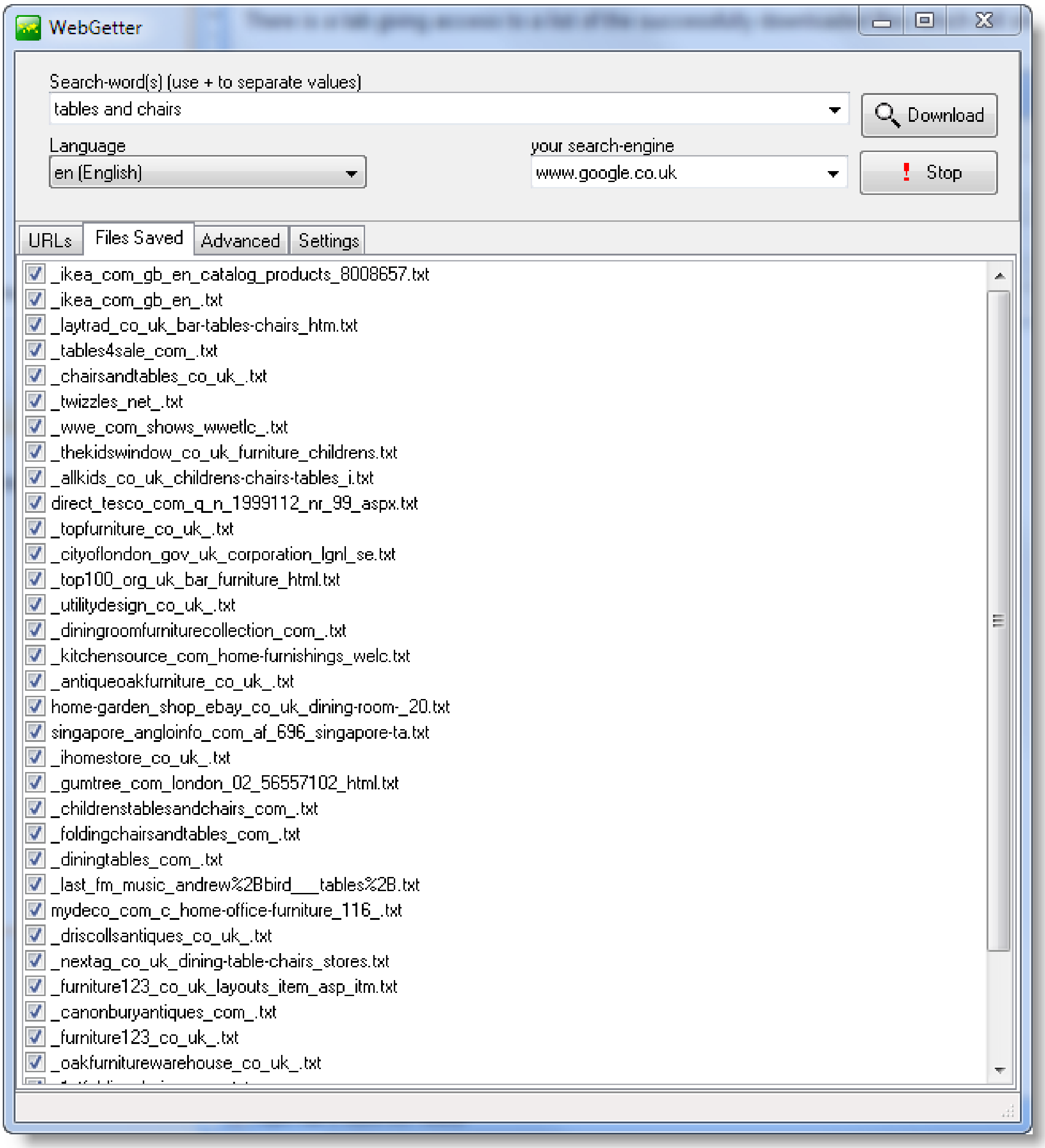
You can specify the whole search URL and terms string yourself if you like with a box in the Advanced options.

See also: Display 326, Limitations 328

10.2.3 display

As **WebGetter** works, it shows the URLs visited. If greyed out, they were too small to be of use or haven't been contacted yet.

There is a tab giving access to a list of the successfully downloaded files which will show something like this.



Search-word(s) (use + to separate values): tables and chairs

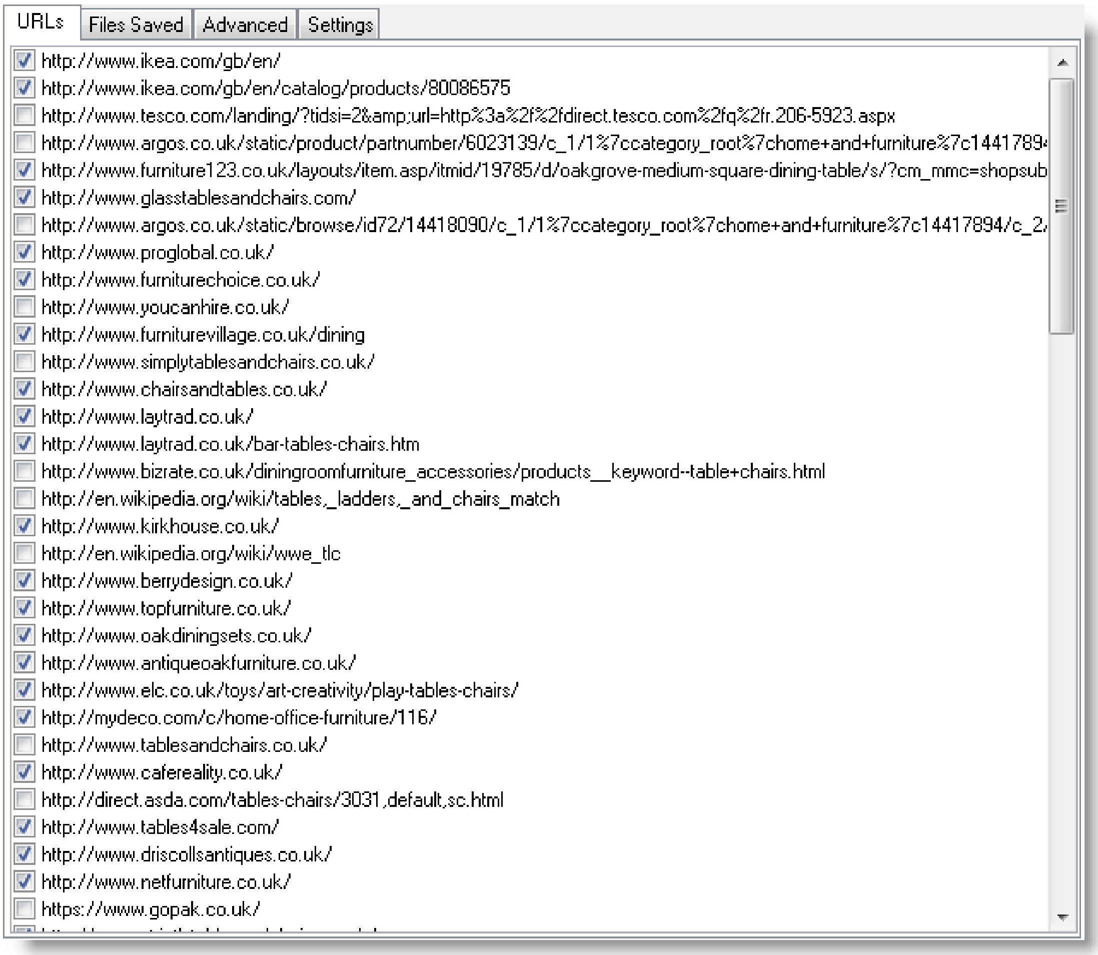
Language: en (English)

your search-engine: www.google.co.uk

☑ _ikea_com_gb_en_catalog_products_8008657.txt
☑ _ikea_com_gb_en_.txt
☑ _laytrad_co_uk_bar-tables-chairs_htm.txt
☑ _tables4sale_com_.txt
☑ _chairsandtables_co_uk_.txt
☑ _twizzles_net_.txt
☑ _wwe_com_shows_wwetlc_.txt
☑ _thekidswindow_co_uk_furniture_childrens.txt
☑ _allkids_co_uk_childrens-chairs-tables_i.txt
☑ direct_tesco_com_q_n_1999112_nr_99_aspx.txt
☑ _topfurniture_co_uk_.txt
☑ _cityoflondon_gov_uk_corporation_lgnl_se.txt
☑ _top100_org_uk_bar_furniture_html.txt
☑ _utilitydesign_co_uk_.txt
☑ _diningroomfurniturecollection_com_.txt
☑ _kitchensource_com_home-furnishings_welc.txt
☑ _antiqueoakfurniture_co_uk_.txt
☑ home-garden_shop_ebay_co_uk_dining-room-_20.txt
☑ singapore_angloinfo_com_af_696_singapore-ta.txt
☑ _ihomestore_co_uk_.txt
☑ _gumtree_com_london_02_56557102_html.txt
☑ _childrenstablesandchairs_com_.txt
☑ _foldingchairsandtables_com_.txt
☑ _diningtables_com_.txt
☑ _last_fm_music_andrew%2Bbird___tables%2B.txt
☑ mydeco_com_c_home-office-furniture_116_.txt
☑ _driscollsantiques_co_uk_.txt
☑ _nextag_co_uk_dining-table-chairs_stores.txt
☑ _furniture123_co_uk_layouts_item_asp_itm.txt
☑ _canonburyantiques_com_.txt
☑ _furniture123_co_uk_.txt
☑ _oakfurniturewarehouse_co_uk_.txt

Double-click a file to view and, if you like, edit it in Notepad.

The URLS list looks like this



☑ http://www.ikea.com/gb/en/
☑ http://www.ikea.com/gb/en/catalog/products/80086575
□ http://www.tesco.com/landing/?tidsi=2&amp;url=http%3a%2f%2fdirect.tesco.com%2fq%2fr.206-5923.aspx
□ http://www.argos.co.uk/static/product/partnumber/6023139/c_1/1%7ccategory_root%7chome+and+furniture%7c1441789
☑ http://www.furniture123.co.uk/layouts/item.asp/itmid/19785/d/oakgrove-medium-square-dining-table/s/?cm_mmc=shopsub
☑ http://www.glasstablesandchairs.com/
□ http://www.argos.co.uk/static/browse/id72/14418090/c_1/1%7ccategory_root%7chome+and+furniture%7c14417894/c_2/
☑ http://www.proglobal.co.uk/
☑ http://www.furniturechoice.co.uk/
□ http://www.youcanhire.co.uk/
☑ http://www.furniturevillage.co.uk/dining
□ http://www.simplytablesandchairs.co.uk/
☑ http://www.chairsandtables.co.uk/
☑ http://www.laytrad.co.uk/
☑ http://www.laytrad.co.uk/bar-tables-chairs.htm
□ http://www.bizrate.co.uk/diningroomfurniture_accessories/products__keyword--table+chairs.html
□ http://en.wikipedia.org/wiki/tables,_ladders,_and_chairs_match
☑ http://www.kirkhouse.co.uk/
□ http://en.wikipedia.org/wiki/wwe_tlc
☑ http://www.berrydesign.co.uk/
☑ http://www.topfurniture.co.uk/
☑ http://www.oakdiningsets.co.uk/
☑ http://www.antiqueoakfurniture.co.uk/
☑ http://www.elc.co.uk/toys/art-creativity/play-tables-chairs/
☑ http://mydeco.com/c/home-office-furniture/116/
□ http://www.tablesandchairs.co.uk/
☑ http://www.cafereality.co.uk/
□ http://direct.asda.com/tables-chairs/3031,default,sc.html
☑ http://www.tables4sale.com/
☑ http://www.driscollsantiques.co.uk/
☑ http://www.netfurniture.co.uk/
□ https://www.gopak.co.uk/

Just double-click an URL to view it in your browser.

See also: Settings 325, Limitations 328

10.2.4 limitations

Everything depends on the search engine and the search terms you use. The Internet is a huge noticeboard; lots of stuff on it is merely ads and catalogue prices etc. The search terms are collected by the search engines by examining terms inserted by the web page author. There is no guarantee that the web pages are really "about" the term you specify, though they should be roughly related in some way.

Use the Settings 325 to be demanding about what you download, e.g. in requiring certain words or phrases to be present.

See also: Display 326

10.3 Corpus Corruption Detector

10.3.1 Aim

The purpose is to check whether one or more of your text files in your corpus doesn't belong. This could be because

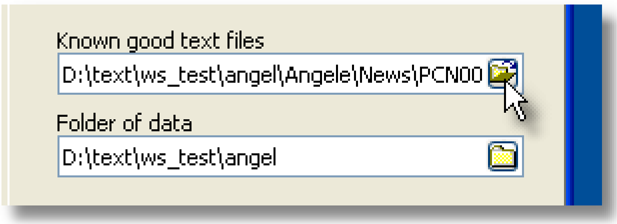
- it has got corrupted so what used to be good text is now just random characters or has got cut much shorter because of disk problems
- it isn't even in the same language as the rest of the corpus

The tool works in any language. It does it by using a known sample of good text (in whatever language) and comparing that good text with all your corpus.

See also : How to do it 329

10.3.2 How it works

1. Choose a set of "known good text files" which you're sure of. The program uses these to evaluate the others.



Known good text files
D:\text\ws_test\angel\Angele\News\PCN00

Folder of data
D:\text\ws_test\angel

When you click the button for known good text files, you can choose a number. You might choose 20 good ones so as to get a lot of information about what your corpus is like.

2. Choose your corpus head folder and check the "include sub-folders" box if your corpus spreads over that folder and sub-folders.

3. The program will anyway look out for oddities such as a text file which has holes 466 in it, eg. where the system thinks it's 1000 characters long but there are only 700.

4. If you check the "digraph check" box it will additionally check that the pairs of letters (digraphs) are of roughly the right frequency in each text file. For example there should be a lot of TH combinations if your text is in English, and no QF combinations. If you are working with a corpus in Portuguese and your text files are in Portuguese too, of course the digraphs will be different, and TH won't be frequent. The program ignores punctuation.

5. If you are doing a digraph check you can vary certain parameters such as how much variation there may be between the frequencies of the digraphs (a sensible setting for "frequency variation per 1000" could be 30 (in other words 3%)), and "percent fail allowed" (which might be set at say 25 -- this means that up to 25% of the digraph pairs may be out of balance before an alert is sounded).

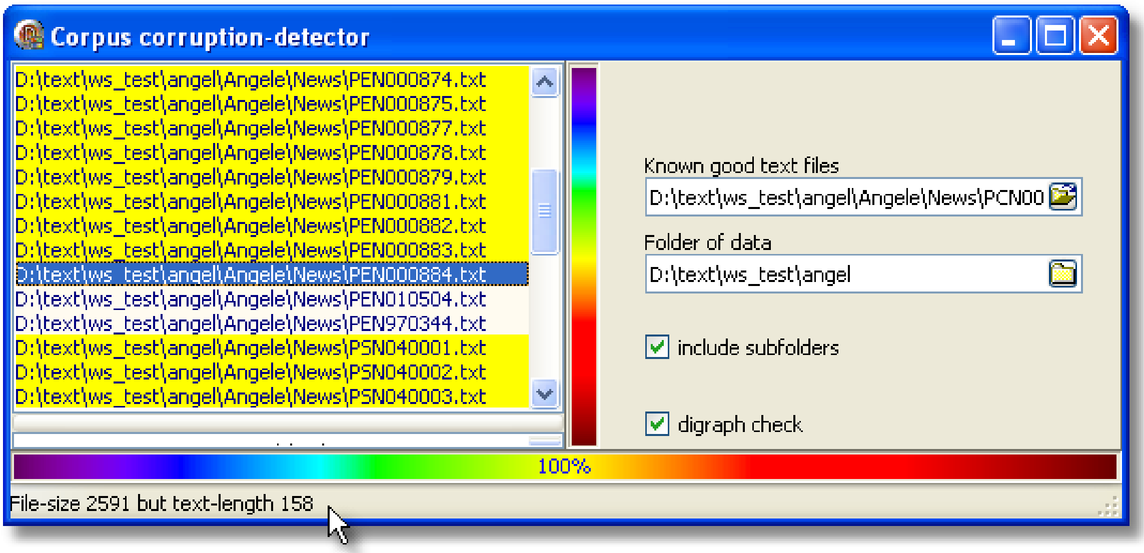
6. Press Start.

You will see the progress bar moving forward.

If you see a file-name in the top-left box, a click on it will indicate why it was found questionable. Double-clicking it will open up the text in the window below so you can examine it carefully.

Filenames of possibly corrupted texts are yellow if the basic check fails, and cream-coloured if the reason is because of a diagraph mis-match.

In the screenshot, PEN000884.txt is problematic because the file-size on disk is 2591 (there should be 2591 characters) but there are only 158, as shown in the statusbar at the bottom.



D:\text\ws_test\angel\Angele\News\PEN000874.txt
D:\text\ws_test\angel\Angele\News\PEN000875.txt
D:\text\ws_test\angel\Angele\News\PEN000877.txt
D:\text\ws_test\angel\Angele\News\PEN000878.txt
D:\text\ws_test\angel\Angele\News\PEN000879.txt
D:\text\ws_test\angel\Angele\News\PEN000881.txt
D:\text\ws_test\angel\Angele\News\PEN000882.txt
D:\text\ws_test\angel\Angele\News\PEN000883.txt
D:\text\ws_test\angel\Angele\News\PEN000884.txt
D:\text\ws_test\angel\Angele\News\PEN010504.txt
D:\text\ws_test\angel\Angele\News\PEN970344.txt
D:\text\ws_test\angel\Angele\News\PSN040001.txt
D:\text\ws_test\angel\Angele\News\PSN040002.txt
D:\text\ws_test\angel\Angele\News\PSN040003.txt

Known good text files: D:\text\ws_test\angel\Angele\News\PCN00

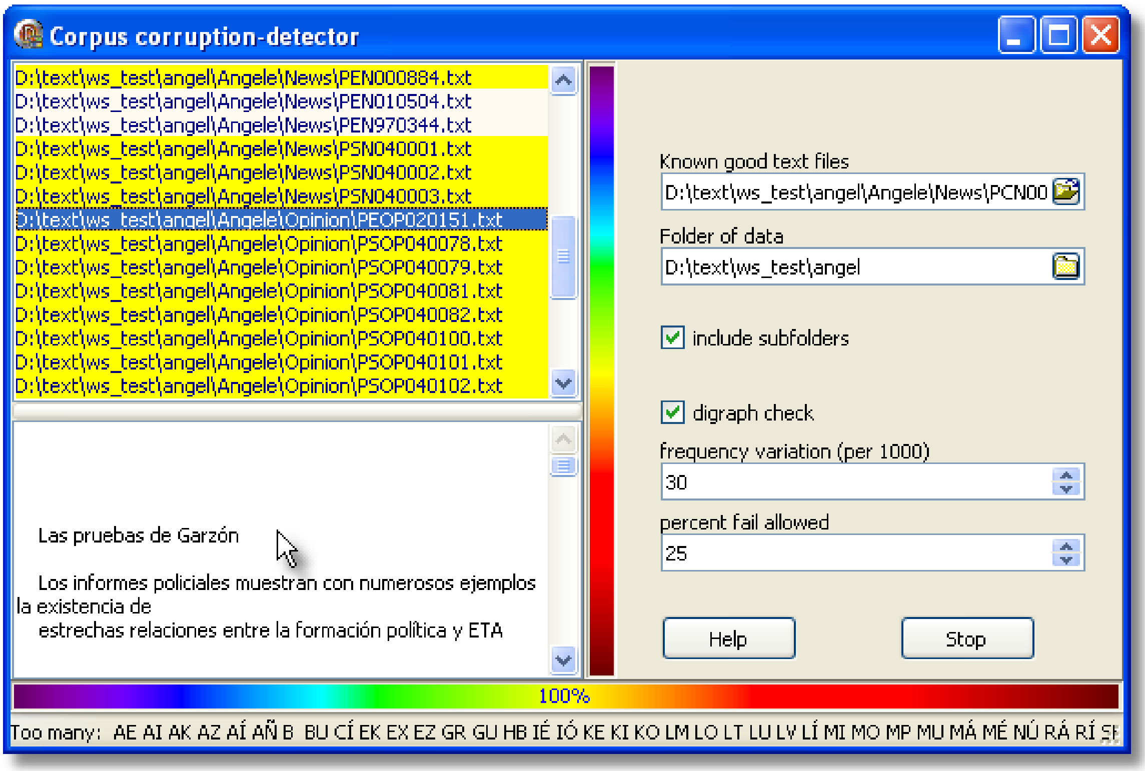
Folder of data: D:\text\ws_test\angel

☑ include subfolders

☑ digraph check

File-size 2591 but text-length 158

In the case of PEOP020151.txt, the text appears below (after double-clicking the list),



D:\text\ws_test\angel\Angele\News\PEN000884.txt
D:\text\ws_test\angel\Angele\News\PEN010504.txt
D:\text\ws_test\angel\Angele\News\PEN970344.txt
D:\text\ws_test\angel\Angele\News\PSN040001.txt
D:\text\ws_test\angel\Angele\News\PSN040002.txt
D:\text\ws_test\angel\Angele\News\PSN040003.txt
D:\text\ws_test\angel\Angele\Opinion\PEOP020151.txt
D:\text\ws_test\angel\Angele\Opinion\PSOP040078.txt
D:\text\ws_test\angel\Angele\Opinion\PSOP040079.txt
D:\text\ws_test\angel\Angele\Opinion\PSOP040081.txt
D:\text\ws_test\angel\Angele\Opinion\PSOP040082.txt
D:\text\ws_test\angel\Angele\Opinion\PSOP040100.txt
D:\text\ws_test\angel\Angele\Opinion\PSOP040101.txt
D:\text\ws_test\angel\Angele\Opinion\PSOP040102.txt

Las pruebas de Garzón

Los informes policiales muestran con numerosos ejemplos
la existencia de
estrechas relaciones entre la formación política y ETA

Known good text files: D:\text\ws_test\angel\Angele\News\PCN00

Folder of data: D:\text\ws_test\angel

☑ include subfolders

☑ digraph check

frequency variation (per 1000): 30

percent fail allowed: 25

Too many: AE AI AK AZ AÍ AÑ B BU CÍ EK EX EZ GR GU HB IÉ IÓ KE KI KO LM LO LT LU LV LÍ MI MO MP MU MÁ MÉ NÚ RÁ RÍ SI

and the status bar says the tool has found an imbalance in the digraphs. The text itself has a lot of blank space at the top but otherwise looks OK (it is supposed to be in Spanish) but the detector has flagged it up as possibly defective.

10.4 Minimal Pairs

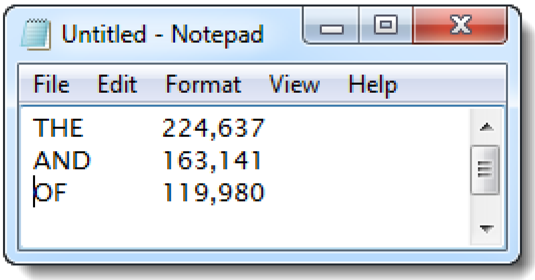
10.4.1 aim

A program for finding possible typos and pairs of words which are minimally different from each other (minimal pairs). For example, you may have a word list which contains ALEADY 5 and ALREADY 461, that is, your texts contain 5 instances where there is a possible misprint and 461 which are correct. This program helps to find possible variants and typos and anagrams.

See also : requirements 332, choosing your files 333, output 336, rules and settings 337, running the program 338.

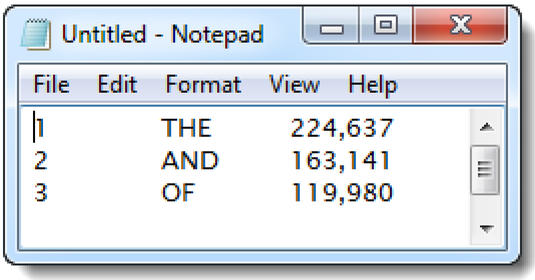
10.4.2 requirements

A word-list in text format. Each line should contain a word and its frequency separated by tabs, e.g.



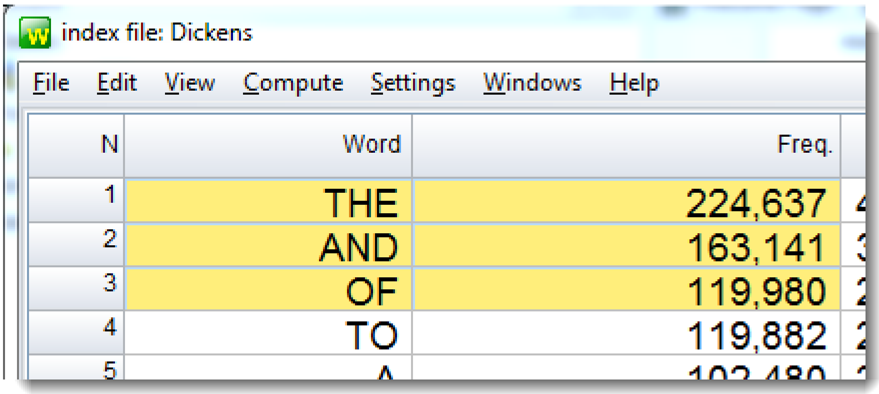
```
THE	224,637
AND	163,141
OF	119,980
```

or



```
1	THE	224,637
2	AND	163,141
3	OF	119,980
```

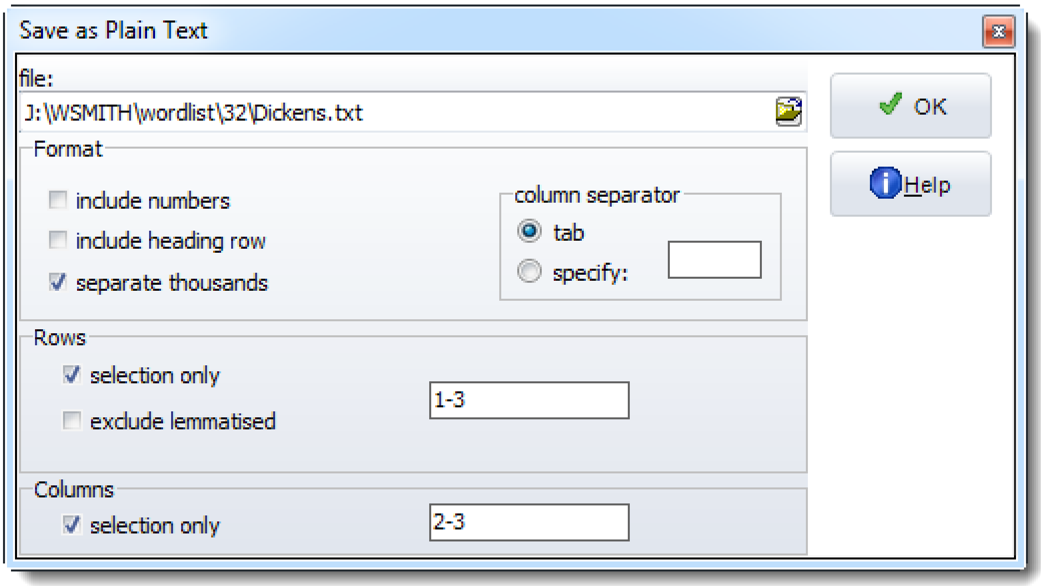
You can make such a list using WordList 6. For example, select (highlight) the columns containing the word and its frequency and



N	Word	Freq.
1	THE	224,637
2	AND	163,141
3	OF	119,980
4	TO	119,882
5	A	102,480

copy to the clipboard, then paste into Notepad, or

save as TXT (without numbers or heading row):



Save as Plain Text

file: J:\WSMITH\wordlist\32\Dickens.txt

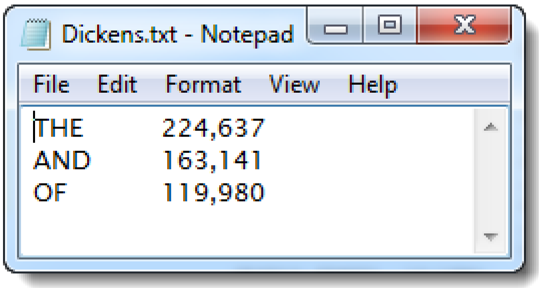
Format
- □ include numbers
- □ include heading row
- ☑ separate thousands

column separator
- ● tab
- ○ specify:

Rows
- ☑ selection only 1-3
- □ exclude lemmatised

Columns
- ☑ selection only 2-3

giving this:



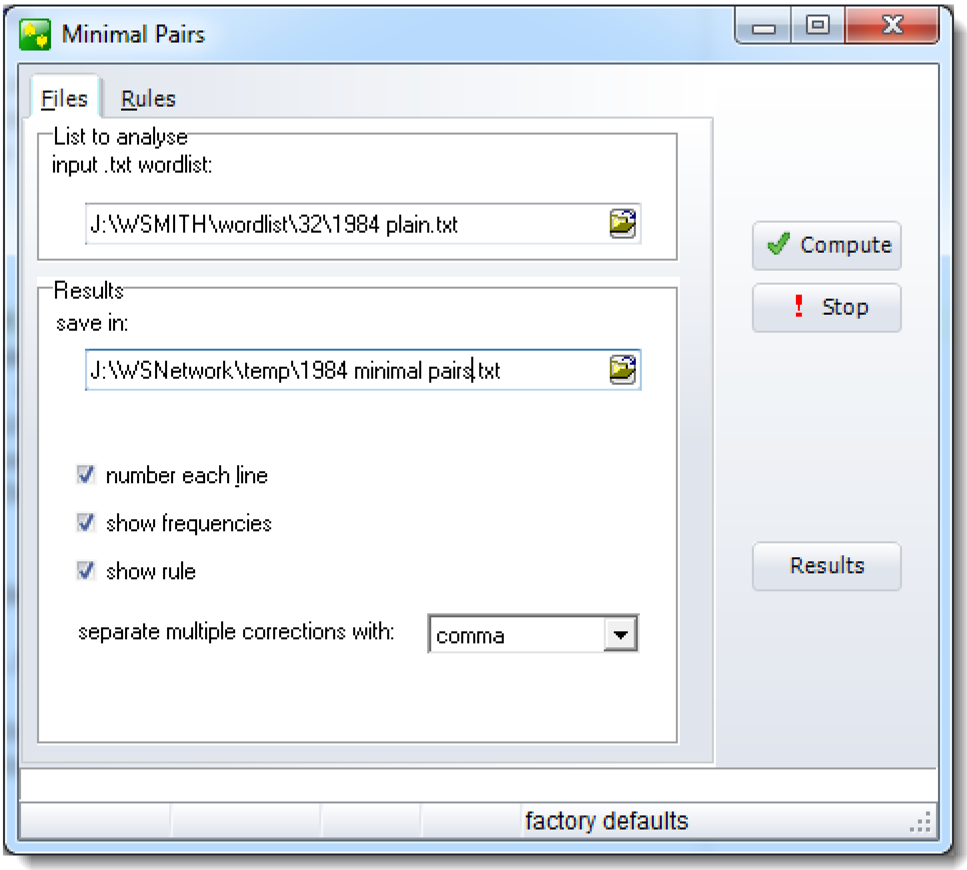
Dickens.txt - Notepad

```
THE     224,637
AND     163,141
OF      119,980
```

See also : aim 331, choosing your files 333, output 336, rules and settings 337, running the program 338 .

10.4.3 choosing your files

Choose your input word list (which must be in plain text format) by clicking the button at the right of the edit space and finding the word list .txt file.



List to analyse

input .txt wordlist:

J:\WSMITH\wordlist\32\1984 plain.txt

Results

save in:

J:\WSNetwork\temp\1984 minimal pairs.txt

☑ number each line

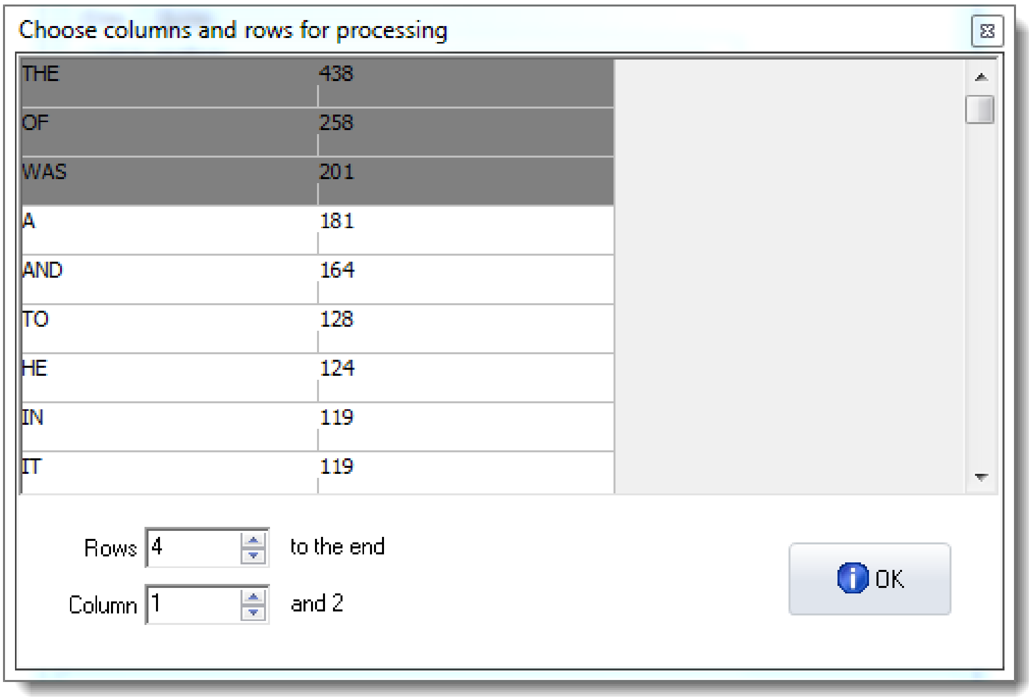
☑ show frequencies

☑ show rule

separate multiple corrections with: comma

Type in an appropriate file-name for your results. Choose the rules 337 too. When you're ready, press the *Compute* button.

You'll then be asked to choose the columns and rows (allowing you to skip header lines or the number column if your txt file has those).



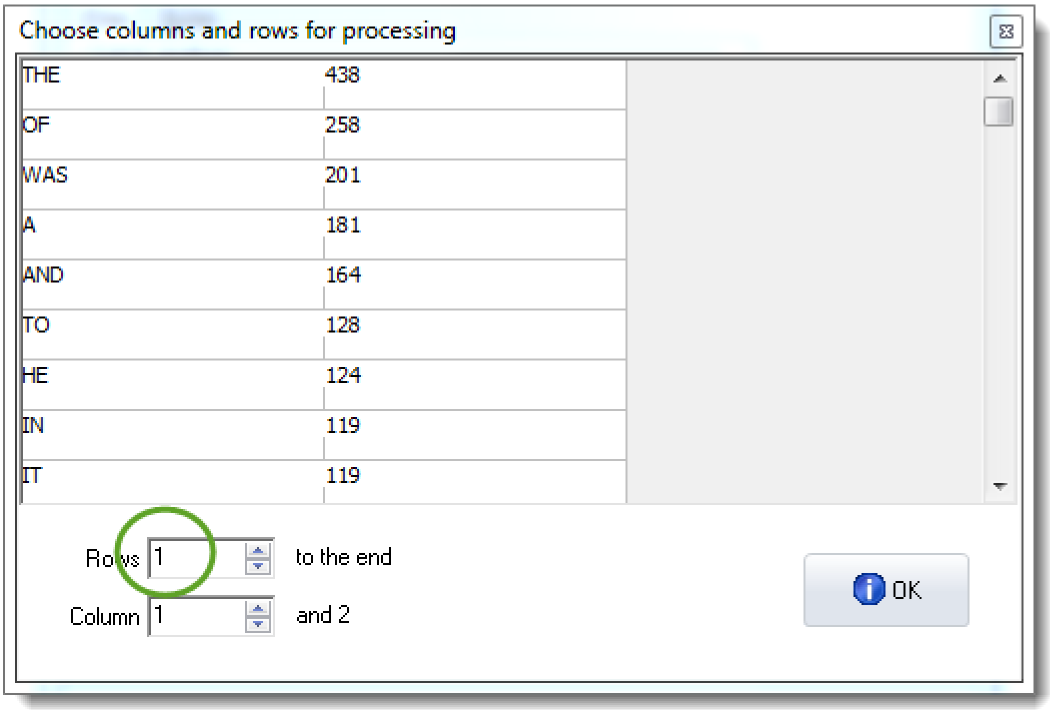
Choose columns and rows for processing

THE	438
OF	258
WAS	201
A	181
AND	164
TO	128
HE	124
IN	119
IT	119

Rows 4 to the end

Column 1 and 2

Here, the first three lines are greyed out, so we need to alter the Rows box:



Choose columns and rows for processing

THE	438
OF	258
WAS	201
A	181
AND	164
TO	128
HE	124
IN	119
IT	119

Rows 1 to the end

Column 1 and 2

See also : aim 331, requirements 332, output 336, rules and settings 337, running the program 338.

10.4.4 output

An example of output is

```
418    ALTHOUGHT     (7)    ALTHOUGH(37975)
```

Here the lines are numbered, and the bracketed numbers mean that ALTHOUGHT occurred 7 times and ALTHOUGH 37,975 times.

An example using Dutch medical text, lower case:

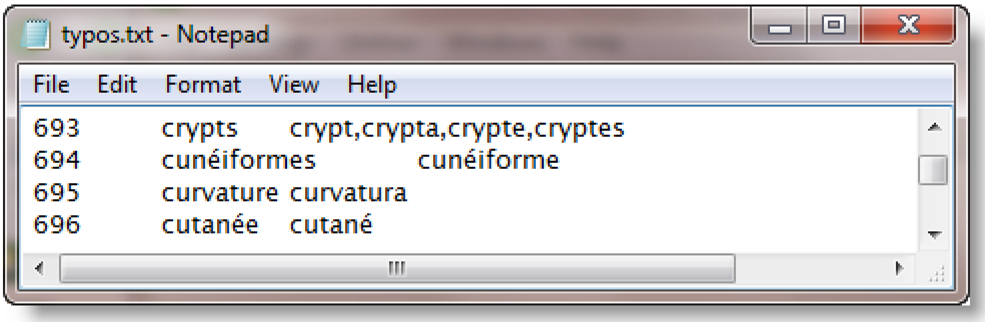
```
136    aplasie       (1)    aplasia(1)[L]
137    apyogene      (1)    apyogeen(1)[S]
138    arachnoideales       (1)    arachnoidales(1)[I]
```

Here line 136 generated a 1-Letter difference, 137 a Swap and 138 an Insertion.

An example using Guardian newspaper, looking for anagrams:

```
35     AUDIE  (7)    ADIEU(43)[A]
36     ABASS  (6)    ASSAB(16)[A]
37     AGUIAR (6)    AURIGA(11)[A]
38     ALRED'S       (6)    ADLER'S(18)[A]
39     ANDOR  (6)    ADORN(128)[A]
```

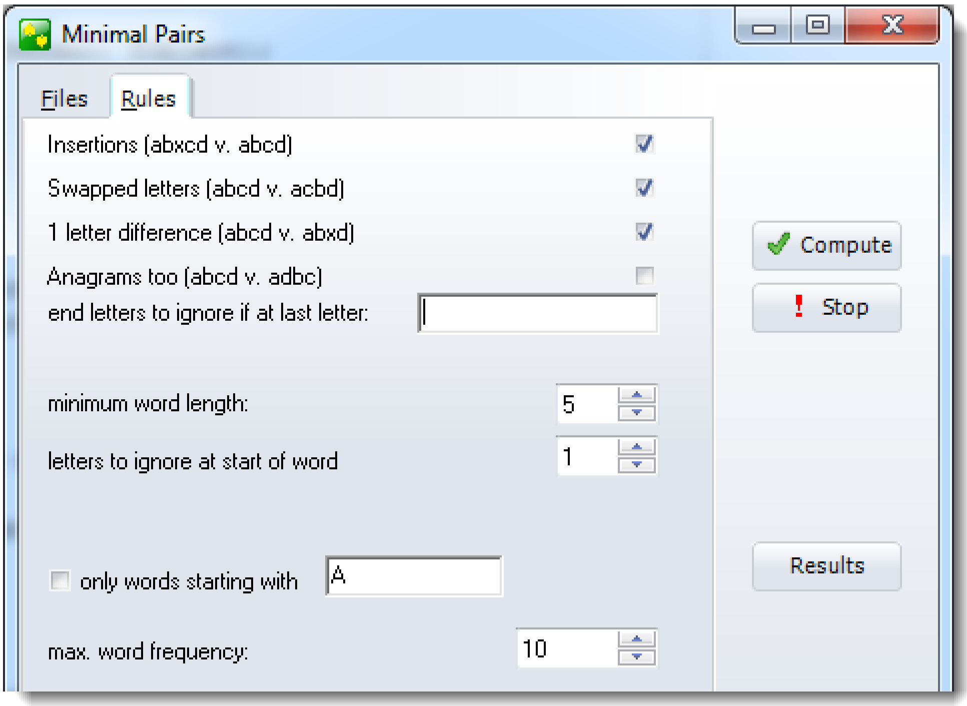
an example where the alternatives are separated with commas but the rule and frequencies are not shown.



```
693      crypts     crypt,crypta,crypte,cryptes
694      cunéiformes        cunéiforme
695      curvature curvatura
696      cutanée   cutané
```

See also : aim 331, requirements 332, choosing your files 333, rules and settings 337, running the program 338.

10.4.5 rules and settings



Minimal Pairs

Files | Rules

Insertions (abxcd v. abcd) ☑

Swapped letters (abcd v. acbd) ☑

1 letter difference (abcd v. abxd) ☑

Anagrams too (abcd v. adbc) □

end letters to ignore if at last letter:

minimum word length: 5

letters to ignore at start of word 1

□ only words starting with A

max. word frequency: 10

Rules

Insertions (abxcd v. abcd)

This rule looks for 1 extra letter which may be inserted, e.g. HOWWEVER

Swapped letters (abcd v. acbd)

This rule looks for letters which have got swapped, e.g. HOVEWER

1 letter difference (abcd v. abxd)

This rule looks for a 1 letter difference, e.g. HOWEXER

Anagrams too (abcd v. adbc)

This rule looks for the same letters in a different order, e.g. HWVROEE

Settings:

end letters to ignore if at last letter:

This rule allows you to specify any letters to ignore if at the end of the word, e.g. if you specify "s", the possibility of a typo when comparing ELEPHANT and ELEPHANTS will not be reported.

minimum word length

This setting specifies the minimum word length for the program to consider the possibility there is a typo. The default is 5, which means 4-letter words will be simply ignored. This is to speed up processing, and because most typos probably occur in longer words.

letters to ignore at start of word

This setting (default =1) allows you to assume that when looking for minimal pairs there is a part of each at the beginning which matches perfectly. For example, when considering ALEADY, the program probably doesn't need to look beyond words beginning with A for minimal pairs. If the setting is 1, it will not find BLEADY as a minimal pair. To check all words, take the setting down to 0. The program will be 26 times slower as a result!

only words starting with …

If you choose this option, the program will ignore the next setting (max. word frequency). Here you can type in a sequence such as F,G,H and if so, the program will take all words beginning F or G or H (whatever their frequency) and look for minimal pairs based on the rules and settings above.

max. word frequency

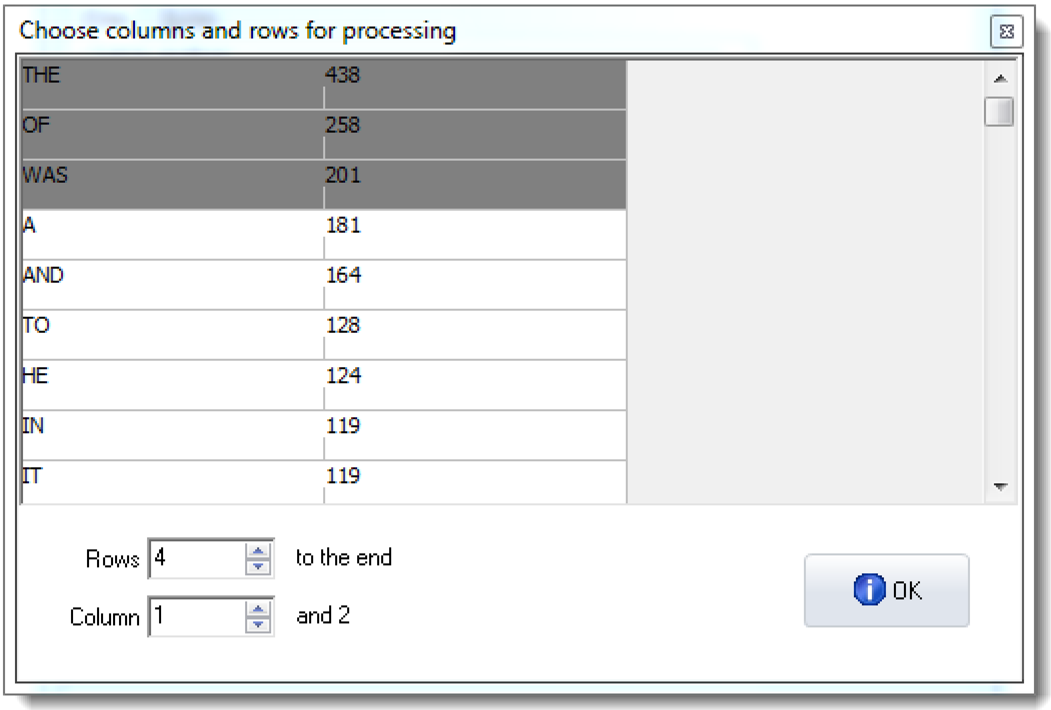
(ignored if "all words starting with" is checked) How frequent can a typo be? This will depend on how much text your word-list is based on. The default is 10, which means that any word which appears 11 times is assumed to be OK, not a typo.

Factory Defaults (restores default values)

See also : aim 331, requirements 332, choosing your files 333, output 336, running the program 338.

10.4.6 running the program

- Press "Compute".
  You should then see your source text, with a few lines visible. Some of the rows and columns may be greyed and others white: move the column and row numbers till the real data are white and any headings or line-numbers are greyed out.



Choose columns and rows for processing

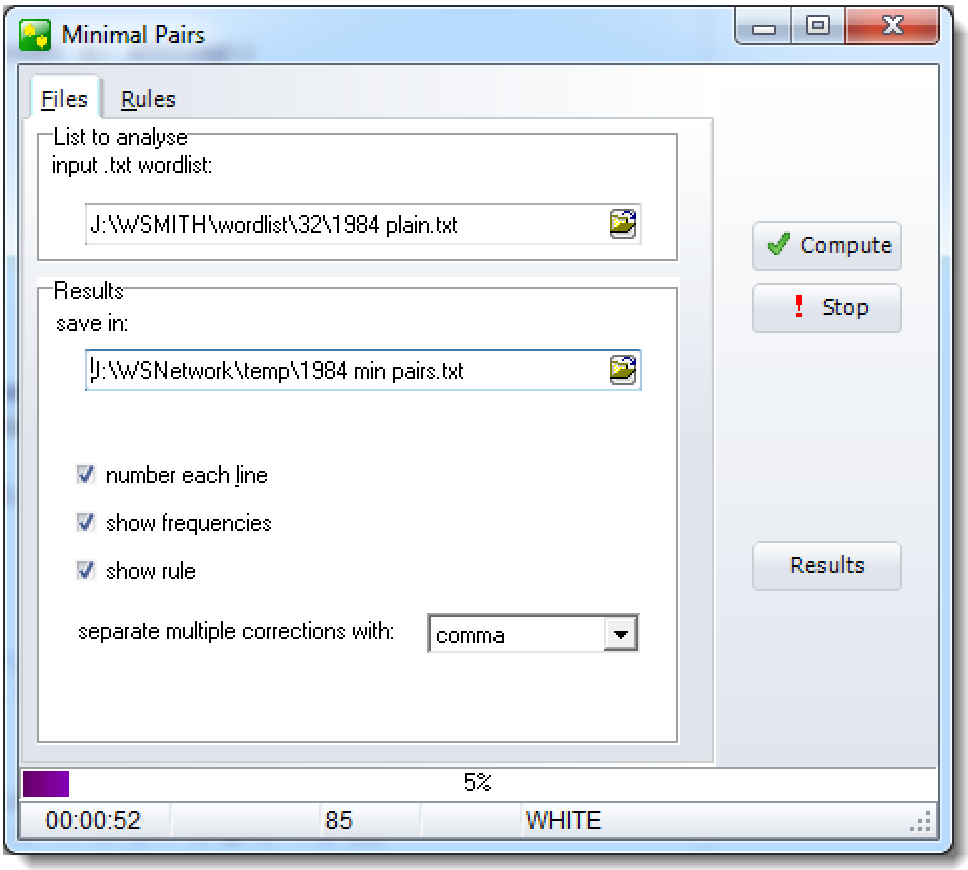
THE	438
OF	258
WAS	201
A	181
AND	164
TO	128
HE	124
IN	119
IT	119

Rows 4 to the end

Column 1 and 2

Here the first three lines are greyed out, and that can be fixed by changing Rows from 4 to 1.

Once you press OK the program starts:



List to analyse
input .txt wordlist:
J:\WSMITH\wordlist\32\1984 plain.txt

Results
save in:
J:\WSNetwork\temp\1984 min pairs.txt

☑ number each line
☑ show frequencies
☑ show rule

separate multiple corrections with: comma

5%
00:00:52 85 WHITE

If you want to stop in the middle, press "Stop".

You can press "Results" to see your results file, when you have finished.

See also : aim 331, requirements 332, choosing your files 333, output 336, rules and settings 337

10.5 File Viewer

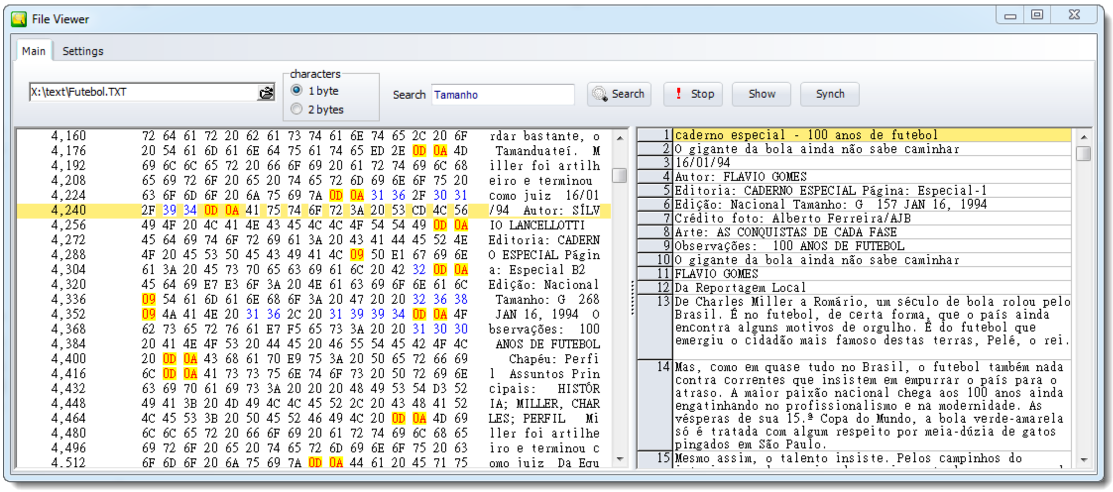
10.5.1 Using File Viewer

Aim

To help you examine files of various kinds to see what is in them. This might be

- in order to see whether they’re really in plain text format
- to see whether there's something wrong with them, such as unusual characters which oughtn't to be there
- to see whether they were formatted for Windows, Mac, or for Unix

- to check out any hidden stuff in there. (A **Word .doc** 444 for example will have lots of hidden stuff you don't see on the screen but is inside the file anyway, such as the name of the author, type of printer being used, etc.)
- to find strings of words in a database, a spreadsheet or even a program file.
- to get certain selected characters picked out in an easy-to-find colour



X:\text\Futebol.TXT

characters
● 1 byte
○ 2 bytes

Search Tamanho

```
   4,160     72 64 61 72 20 62 61 73 74 61 6E 74 65 2C 20 6F   rdar bastante, o
   4,176     20 54 61 6D 61 6E 64 75 61 74 65 ED 2E 0D 0A 4D    Tamanduateí.  M
   4,192     69 6C 6C 65 72 20 66 6F 69 20 61 72 74 69 6C 68   iller foi artilh
   4,208     65 69 72 6F 20 65 20 74 65 72 6D 69 6E 6F 75 20   eiro e terminou 
   4,224     63 6F 6D 6F 20 6A 75 69 7A 0D 0A 31 36 2F 30 31   como juiz  16/01
   4,240     2F 39 34 0D 0A 41 75 74 6F 72 3A 20 53 CD 4C 56   /94  Autor: SÍLV
   4,256     49 4F 20 4C 41 4E 43 45 4C 4C 4F 54 54 49 0D 0A   IO LANCELLOTTI  
   4,272     45 64 69 74 6F 72 69 61 3A 20 43 41 44 45 52 4E   Editoria: CADERN
   4,288     4F 20 45 53 50 45 43 49 41 4C 09 50 E1 67 69 6E   O ESPECIAL Págin
   4,304     61 3A 20 45 73 70 65 63 69 61 6C 20 42 32 0D 0A   a: Especial B2  
   4,320     45 64 69 E7 E3 6F 3A 20 4E 61 63 69 6F 6E 61 6C   Edição: Nacional
   4,336     09 54 61 6D 61 6E 68 6F 3A 20 47 20 20 32 36 38    Tamanho: G  268
   4,352     09 4A 41 4E 20 31 36 2C 20 31 39 39 34 0D 0A 4F    JAN 16, 1994  O
   4,368     62 73 65 72 76 61 E7 F5 65 73 3A 20 20 31 30 30   bservações:  100
   4,384     20 41 4E 4F 53 20 44 45 20 46 55 54 45 42 4F 4C    ANOS DE FUTEBOL
   4,400     20 0D 0A 43 68 61 70 E9 75 3A 20 50 65 72 66 69      Chapéu: Perfi
   4,416     6C 0D 0A 41 73 73 75 6E 74 6F 73 20 50 72 69 6E   l  Assuntos Prin
   4,432     63 69 70 61 69 73 3A 20 20 20 48 49 53 54 D3 52   cipais:   HISTÓR
   4,448     49 41 3B 20 4D 49 4C 4C 45 52 2C 20 43 48 41 52   IA; MILLER, CHAR
   4,464     4C 45 53 3B 20 50 45 52 46 49 4C 20 0D 0A 4D 69   LES; PERFIL   Mi
   4,480     6C 6C 65 72 20 66 6F 69 20 61 72 74 69 6C 68 65   ller foi artilhe
   4,496     69 72 6F 20 65 20 74 65 72 6D 69 6E 6F 75 20 63   iro e terminou c
   4.512     6F 6D 6F 20 6A 75 69 7A 0D 0A 44 61 20 45 71 75   omo juiz  Da Equ
```

```
 1 caderno especial - 100 anos de futebol
 2 O gigante da bola ainda não sabe caminhar
 3 16/01/94
 4 Autor: FLAVIO GOMES
 5 Editoria: CADERNO ESPECIAL Página: Especial-1
 6 Edição: Nacional Tamanho: G  157 JAN 16, 1994
 7 Crédito foto: Alberto Ferreira/AJB
 8 Arte: AS CONQUISTAS DE CADA FASE
 9 Observações:  100 ANOS DE FUTEBOL
10 O gigante da bola ainda não sabe caminhar
11 FLAVIO GOMES
12 Da Reportagem Local
13 De Charles Miller a Romário, um século de bola rolou pelo
   Brasil. É no futebol, de certa forma, que o país ainda
   encontra alguns motivos de orgulho. É do futebol que
   emergiu o cidadão mais famoso destas terras, Pelé, o rei.
14 Mas, como em quase tudo no Brasil, o futebol também nada
   contra correntes que insistem em empurrar o país para o
   atraso. A maior paixão nacional chega aos 100 anos ainda
   engatinhando no profissionalismo e na modernidade. As
   vésperas de sua 15.ª Copa do Mundo, a bola verde-amarela
   só é tratada com algum respeito por meia-dúzia de gatos
   pingados em São Paulo.
15 Mesmo assim, o talento insiste. Pelos campinhos do
```

Here you can see the gory details of the text. Some characters are highlighted in different colours so you can see exactly how the text is formatted.

Loading a text file

Choose your file – if necessary click on the button at the right of the text-input box. Press *Show.*

Characters

The two options available are as 1 bytes or 2 to represent each character-symbol in the text in question. You may need to alter this setting to see your text in a readable format.

The two windows

The left window shows how the "text" is built up. You can see each character as a number and, further to the right, as a character.

The right window shows the text, paragraph by paragraph, word-wrapped so you can read it.

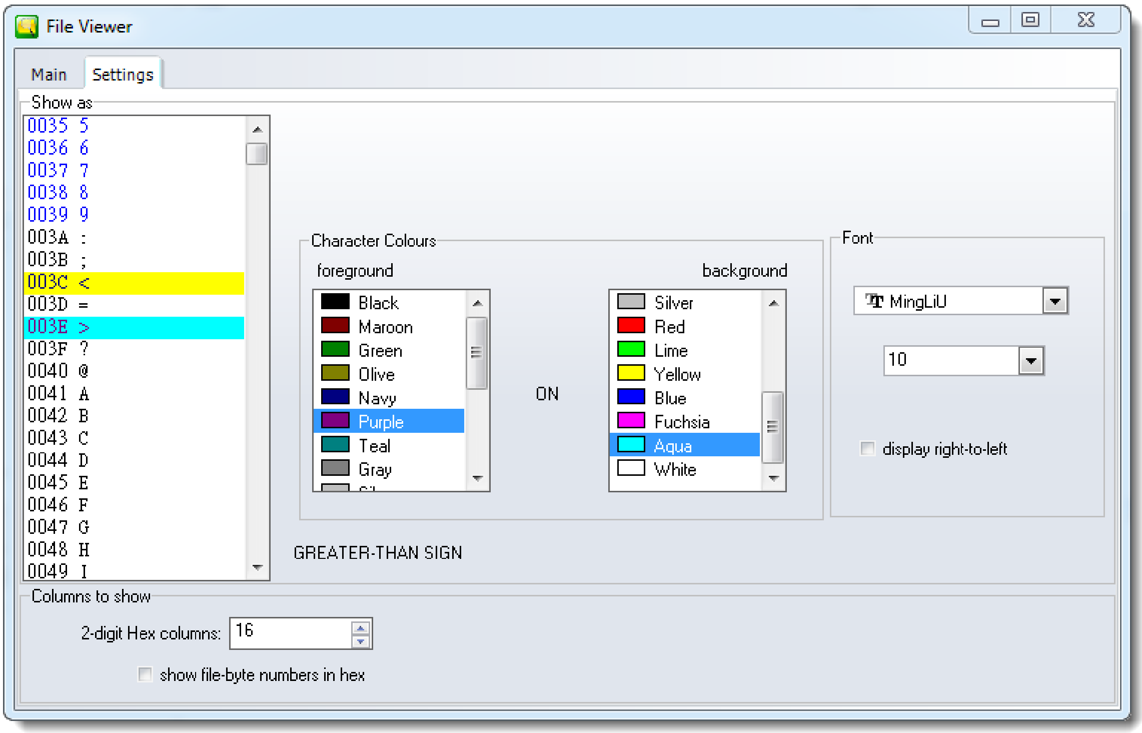
Searching

Just type in the search-word and press Search. The search is case sensitive and is not a "whole word" search.

Synch

Press the *Synch* button to synchronise the two displays. The display you clicked last is the "boss" for synchronising.

Settings



```
0035 5
0036 6
0037 7
0038 8
0039 9
003A :
003B ;
003C <
003D =
003E >
003F ?
0040 @
0041 A
0042 B
0043 C
0044 D
0045 E
0046 F
0047 G
0048 H
0049 I
```

Purple ON Aqua

GREATER-THAN SIGN

Font: MingLiU, 10

☐ display right-to-left

2-digit Hex columns: 16

☐ show file-byte numbers in hex

Colours

The colour grids let you see the number section in special colours, so you can find the potential problems you're interested in.

- First select the character you want coloured.
- Click the foreground or background colour list change the colour.

The character names are Unicode names. In the picture the symbol with the 003E code (>) is the last one clicked.

Font

Choose the font and size in the font window. You may need to change font if you want to see Chinese etc. represented correctly.

Columns

- You can set the "hex" columns between 2 and 16.
- You can see the numbers at the left of the main window in hex or decimal.

10.6 File Utilities

10.6.1 index

This sub-program supplies a few file utilities for general use:

Compare Two Files 347
File Chunker 348
Find Duplicates 348
Rename 349
Find Holes: for "holes 466" in text files
Splitter 343
Joiner 346
Move files to sub-folder

10.6.2 Splitter

10.6.2.1 Splitter: index

Explanations

What is the Splitter utility and what's it for? 343
Filenames 344
Wildcards 345

See also : WordSmith Main Index 2

10.6.2.2 aim of Splitter

This is a sub-program for splitting large files into lots of small ones. **Splitter** needs to know:

Start/End of Section Separator

The symbol which will act as a start or end-of-text separator: eg. **[FF]** or **<end of story>** or **</Text>** or **!#** or **[FF*]** or **[FF?????]** or **CHAPTER #**

Restrictions:

1 The start/end-of-text marker must occur at the beginning of a line in the original large file.
2 It is case sensitive: **</Text>** will not find **</text>**.
3 The first character in the separator may not be a wildcard 345 such as **#**,***** or **?**.
4 ***** and **#** may occur only once each in the separator.

Splitter will create a new file every time it encounters the start/end-of-text marker you've specified. The end of text box determines whether the line containing the separator gets included in the previous or new text file.

Destination Folder

Where you want the small files to be copied to. (You'll need write permission to access it if on a network.)

Required sizes

The minimum and maximum number of lines that your small files can have (default = 5 and

30,000). Only files within these limits will be saved. This feature is useful for extracting files from very large CD-ROM files.
A "line" means from one <Enter> to the next.

Bracket first line

Whether or not you want the first line of each new text file to be bracketed inside **< >** marks. (If your separator is a start-of-section separator like CHAPTER with a number, you may wish that to be in brackets. And often the first line after an end-of-text symbol will contain some kind of header.) If you don't want it to insert < and > around the line, leave this box unchecked.

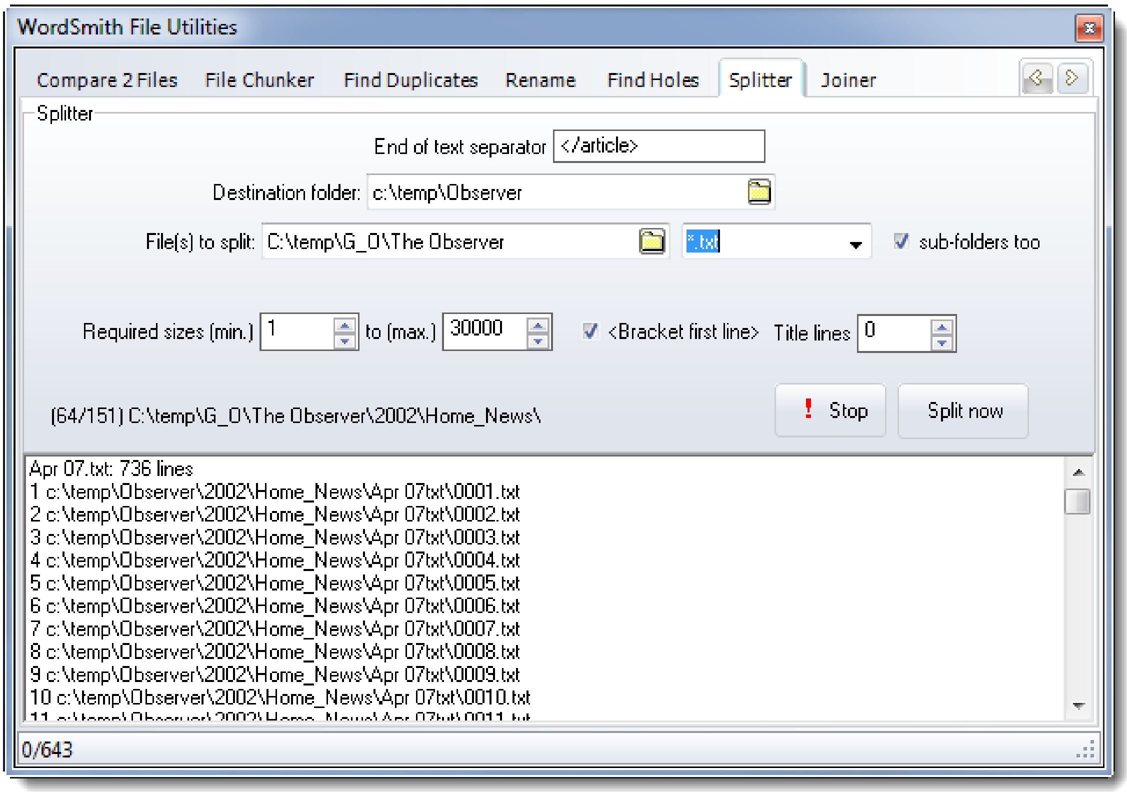
Title Line

If you know that a given line of your texts always contains the title for the sub-text in question, set this counter to that number, otherwise leave it at 0. For example, where you know that every line immediately following **<end of story>** has a title for the next story, you could put 1.
Example :
...
<end of story>
Visiting New York
...
The file-name created for each story will contain the title as well as a suitable number. In this example a file-name might end up as **`C:\texts\split\Visiting New York 0004.txt`**.

See also: Joiner 346, Filenames 344, Wildcards 345, The buttons 441, Text Converter index 355.

10.6.2.3 Splitter: filenames

Splitter will create lots of small files based on your large one(s).



Splitter

End of text separator: </article>

Destination folder: c:\temp\Observer

File(s) to split: C:\temp\G_O\The Observer; *.txt; ☑ sub-folders too

Required sizes (min.): 1; to (max.): 30000; ☑ <Bracket first line>; Title lines: 0

(64/151) C:\temp\G_O\The Observer\2002\Home_News\

```
Apr 07.txt: 736 lines
1 c:\temp\Observer\2002\Home_News\Apr 07txt\0001.txt
2 c:\temp\Observer\2002\Home_News\Apr 07txt\0002.txt
3 c:\temp\Observer\2002\Home_News\Apr 07txt\0003.txt
4 c:\temp\Observer\2002\Home_News\Apr 07txt\0004.txt
5 c:\temp\Observer\2002\Home_News\Apr 07txt\0005.txt
6 c:\temp\Observer\2002\Home_News\Apr 07txt\0006.txt
7 c:\temp\Observer\2002\Home_News\Apr 07txt\0007.txt
8 c:\temp\Observer\2002\Home_News\Apr 07txt\0008.txt
9 c:\temp\Observer\2002\Home_News\Apr 07txt\0009.txt
10 c:\temp\Observer\2002\Home_News\Apr 07txt\0010.txt
```

0/643

It creates filenames as sub-files of a folder based on the name of each text file. In this screenshot, it has found a file called **`C:\temp\G_O\The Observer\2002\Home_news\Apr 07.txt`** and is creating a set of results listed 1 to 11 or more, using the specified destination folder plus the same folder structure as the original texts. Each sub-text is numbered **`0001.txt, 0002.txt`** etc.

Sub-folders are created if there are too many files for a folder.

If a title is detected, each file will contain the title plus a number and .txt. If there is no title, the filename will be the number + .txt added as a file extension.

Tips

1. Splitter will start numbering at 1 each session.
2. Note that the small files will probably take up a lot more room than the original large file did. This is because the disk operating system has a fixed minimum file size. A one-character text file will require this minimum size, which will probably be several thousand bytes in size. Even so, I suggest you keep your text files such that each file is a separate text, by using Splitter. When doing word lists and key words lists, though, do them in batches 39.
3. CD-ROM files when copied to your hard disk may be read-only. You can change this attribute using Text Converter 355.

10.6.2.4 Splitter: wildcards

# The hash symbol, #, is used as a wildcard to represent any ***number***, so **[FF#]** would find *[FF3]* or *[FF9987]* but not *[FF]* or *[FF 9]* (because there's a space in it) or *[FFhello]*.

* The asterisk represents any ***string***, so **[FF*** would find all of the above. * is used as the last character in the end-of-text symbol. It would find *[FF anything at all up to the next <Enter>*.

^ The ^ mark represents any single ***letter***, so **[FF^^]** would find *[FFZQ]* but none of the others.

? The question mark represents any single ***character*** (including spaces, punctuation, letters), so **[FF??]** would find *[FF 9]* in the above examples, but none of the others.

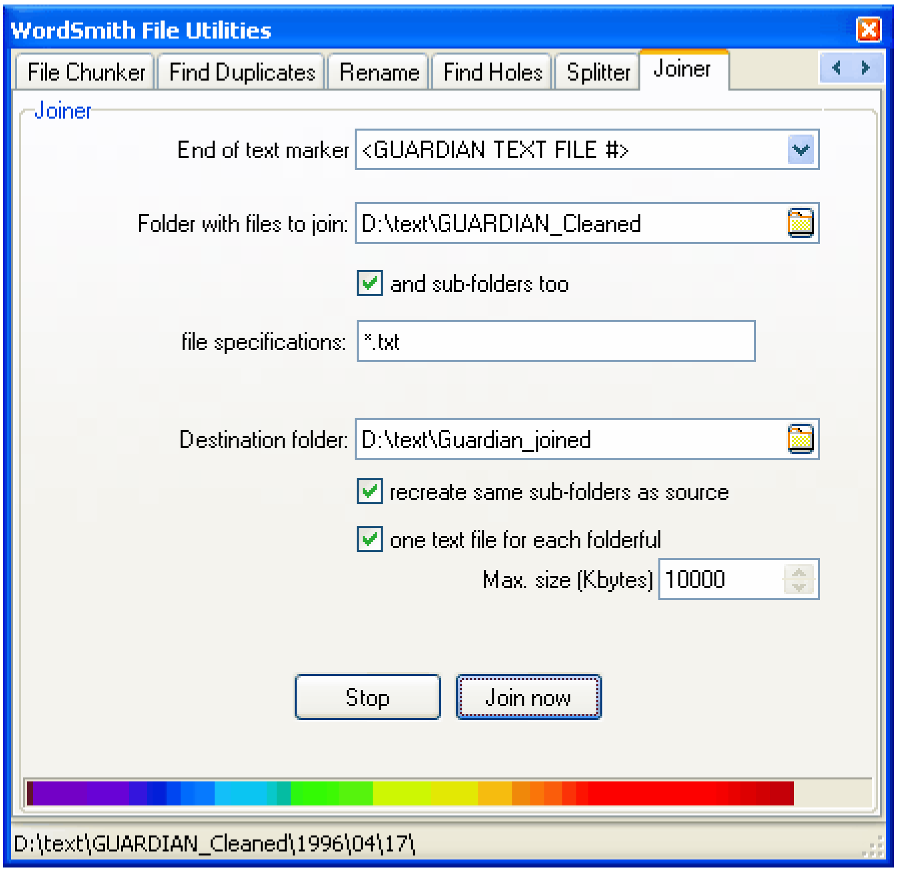
To represent a genuine **#,^,?** or *****, put each one in double quotes, eg. **"?" "#" "^" "*"**.

See also: Settings 343, Wildcards 28

10.6.3 join text files

This is a sub-program for joining small text files into bigger ones. You might want this because you aren't interested in the different texts individually but are only interested in studying the patterns of a whole lot of texts.

When you choose **Joiner** you will see something like this:



End of text marker: <GUARDIAN TEXT FILE #>

Folder with files to join: D:\text\GUARDIAN_Cleaned

☑ and sub-folders too

file specifications: *.txt

Destination folder: D:\text\Guardian_joined

☑ recreate same sub-folders as source

☑ one text file for each folderful

Max. size (Kbytes) 10000

D:\text\GUARDIAN_Cleaned\1996\04\17\

End of text marker

The symbol which will act as an end-of-text separator: eg. **[FF]** or **<end of story>** or **</Text>** or **!#** or **[FF*]** or **[FF?????].** The end-of-text marker will come at the beginning of a line in the original large file. If it includes # this will be replaced by the number of the text as the texts are processed.

Folder with files to join

Where the small files you want to be merged are now. They will not get deleted -- you must merge them into the Destination folder.

and sub-folders too

Check this if you want to process sub-folders of the "folder with files to join".

file specifications

The kinds of text files you want to merge, eg. **`*.*`** or **`*.txt`** or **`*.txt;*.ctx`**.

Destination Folder

Where you want the small files to be copied and merged to. (You'll need write permission to access it if on a network.)

recreate same sub-folders as source

If checked, creates the same structure as in the source. In the example, all the sub-folders of **`d:\text\guardian_cleaned`** will be created below **`d:\text\guardian_joined`**.

one text for each folderful

if checked, a whole folderful of source texts will go into one text file in the destination.

Max. size (Kbytes)

The maximum size in kilobytes that you want the each merged text file to be. 1000 means you will get almost 1 megabyte of text into each. That is about 150,000 words if there are no tags and the text is in English. This only applies if *one text for each folderful* isn't checked.

Stop button

Does what it says on the caption.

See also: Splitter 343, Text Converter index 355.

10.6.4 compare two files

The point of it

The idea is to be able to check whether 2 files are similar or not. You may often make copies of files and a few weeks later cannot remember what they were. Or you have used File Chunker 348 to copy a big file to floppies and want to be sure the copy is identical to the original.

This program checks whether

a) they are the same size

b) they have the same contents

(it goes through both, byte by byte, checking whether they match)

c) they have the same attributes

(file attributes can be "read only" [you cannot alter the file], "system" [a file which Windows thinks is central to your operating system], "hidden" [one which is so important that Bill Gates may be reluctant to even let you know it exists on your disk])

d) they have the same time & date.

How to do it

Specify your 2 files and simply press "Compare".

See also : file chunker 348, find duplicates 348, rename 349

10.6.5 file chunker

The point of it

The idea is to be able to cut up a big file into pieces, so that you can copy it in chunks e.g. for emailing.

Naturally, you may later want to restore the chunks to one file.

How to do it: to copy a file

1. Specify your "file to chunk" (the big one you want to copy)
2. Specify your "drive & folder" (where you want to copy the chunks to.
3. Specify the "size of each chunk"
4. Specify whether to "compress while chunking" (compresses the file as it goes along)
5. Press "Copy".

How to do it: to restore a file

1. Specify your "first chunk" (the first chunk you made using this program)
2. Specify which folder to "restore to" (where you want the results)
3. Specify whether to "delete chunks afterwards" (if they are not needed)
4. Press "Restore".

See also : compare two files 347, find duplicates 348, rename 349

10.6.6 find duplicates

The point of it

The idea is to be able to check whether you have files with the same name in different folders. You may often make copies of files and a few weeks later cannot remember where they were.

By default this program only checks whether the files it is comparing have the same name but dates and file-size can be compared too. It handles lots of folders, the point being to locate unnecessarily duplicated files or confusing reuse of the same filenames.

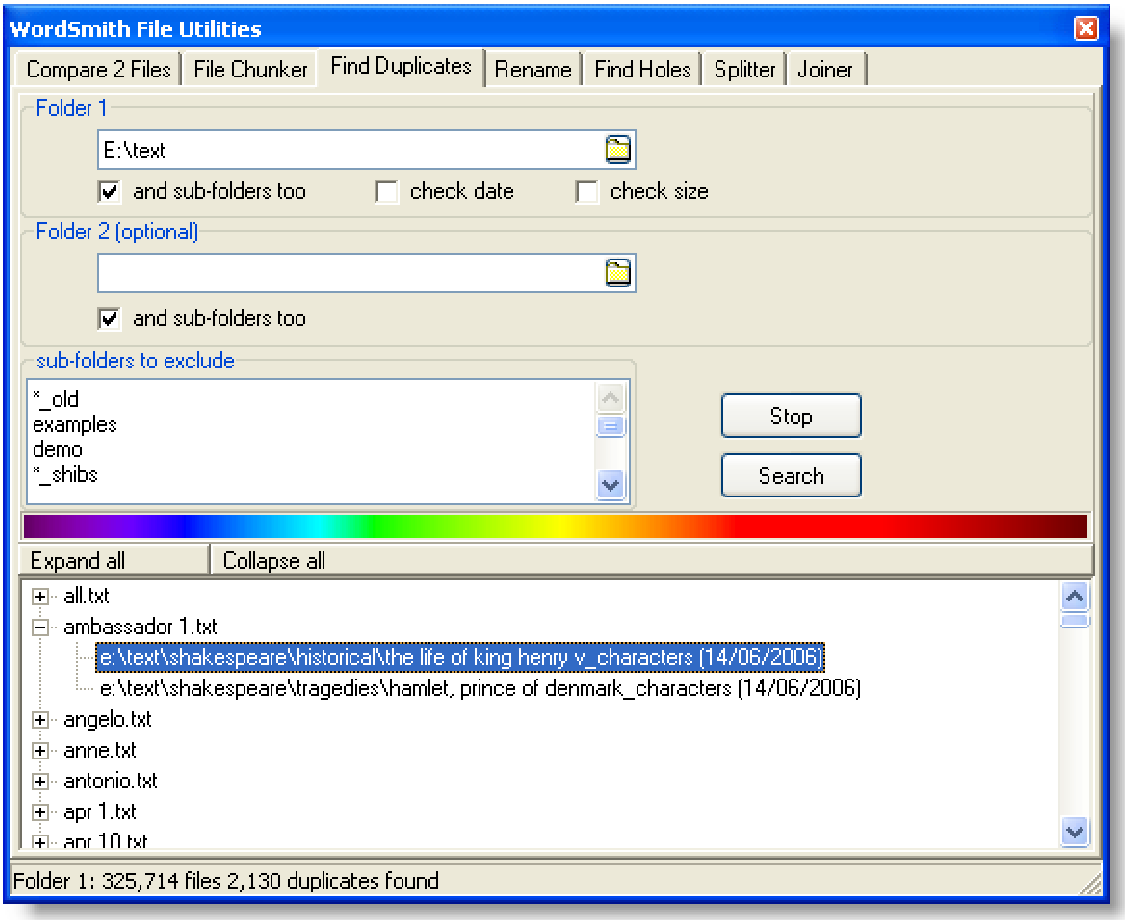
How to do it

Specify your Folder 1 and simply press "Search". Find Duplicates will go through that folder and any sub-folders and will report any duplicates found.

Or you can specify 2 different folders (e.g. on different drives) and the process compares one set with the other.

Sub-folders to exclude

Useful if there are some sub-folders you know you're not interested in. In the example below, any folder whose name ends **`_old`** or **`_shibs`** or whose name is **`demo`** or **`examples`** will be ignored as will any sub-folder below it.



Folder 1

E:\text

☑ and sub-folders too □ check date □ check size

Folder 2 (optional)

☑ and sub-folders too

sub-folders to exclude

*_old
examples
demo
*_shibs

```
⊞ all.txt
⊟ ambassador 1.txt
    e:\text\shakespeare\historical\the life of king henry v_characters (14/06/2006)
    e:\text\shakespeare\tragedies\hamlet, prince of denmark_characters (14/06/2006)
⊞ angelo.txt
⊞ anne.txt
⊞ antonio.txt
⊞ apr 1.txt
⊞ apr 10.txt
```

Folder 1: 325,714 files 2,130 duplicates found

In the window below, you will find all the duplicates listed with the folder and date. In the example we can see there are two files called **`ambassador 1.txt`** in different shakespeare folders.

See also : compare two files 347, file chunker 348, rename 349

10.6.7 rename

The point of it

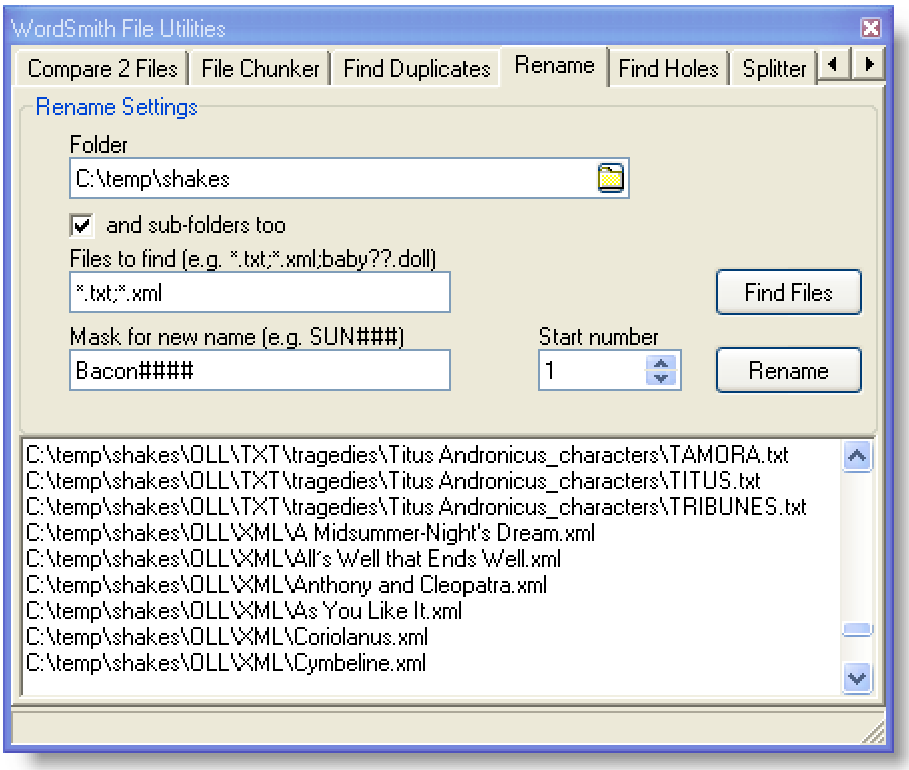
To rename a lot of files at once, in one or more folders. You may have files with excessively long names which do not suit certain applications. Or it is a pain to rename a lot of files one by one.

The idea is to rename a set of files with a standard name plus a number. For example suppose you have downloaded a lot of text files containing emails from the Enron scandal, you could rename them

```
Enron001.txt
Enron002.txt
```

etc.

How to do it



Folder: C:\temp\shakes

☑ and sub-folders too

Files to find (e.g. *.txt;*.xml;baby??.doll): *.txt;*.xml

Mask for new name (e.g. SUN###): Bacon####

Start number: 1

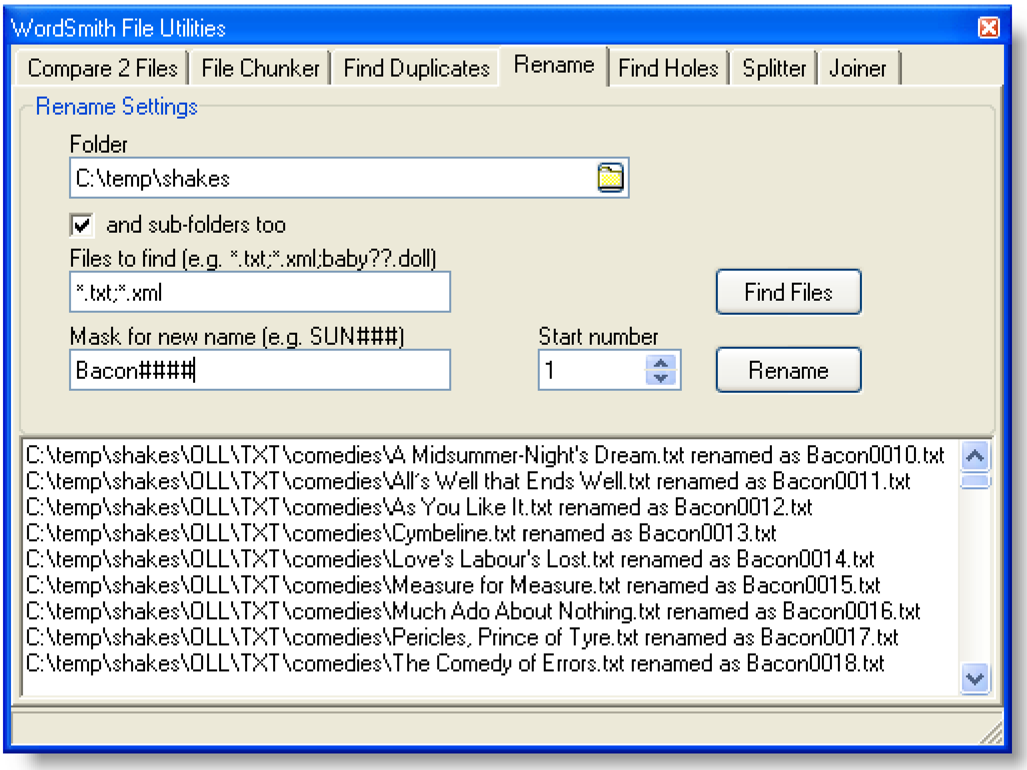
```
C:\temp\shakes\OLL\TXT\tragedies\Titus Andronicus_characters\TAMORA.txt
C:\temp\shakes\OLL\TXT\tragedies\Titus Andronicus_characters\TITUS.txt
C:\temp\shakes\OLL\TXT\tragedies\Titus Andronicus_characters\TRIBUNES.txt
C:\temp\shakes\OLL\XML\A Midsummer-Night's Dream.xml
C:\temp\shakes\OLL\XML\All's Well that Ends Well.xml
C:\temp\shakes\OLL\XML\Anthony and Cleopatra.xml
C:\temp\shakes\OLL\XML\As You Like It.xml
C:\temp\shakes\OLL\XML\Coriolanus.xml
C:\temp\shakes\OLL\XML\Cymbeline.xml
```

Specify your Folder, whether sub-folders will also be processed, and the kinds of file you want to find for renaming.

In the screenshot, **`*.txt;*.xml`** has been specified, which means all **`.txt`** files and all **`.xml`** files. *Find Files* has been pressed, too. In the list you can see some of each.

If you typed **`baby??.doll`** you'd get all files with the **`.doll`** ending as long as the first 4 characters were **`baby`** as in **`baby05.doll, babyyz.doll`**, etc.

Now specify a "mask for new name" and a starting number. The mask can end with a series of # characters standing for numbers. In this screenshot, there are 4 # symbols



Rename Settings

Folder: C:\temp\shakes

☑ and sub-folders too

Files to find (e.g. *.txt;*.xml;baby??.doll): *.txt;*.xml

Mask for new name (e.g. SUN###): Bacon####

Start number: 1

```
C:\temp\shakes\OLL\TXT\comedies\A Midsummer-Night's Dream.txt renamed as Bacon0010.txt
C:\temp\shakes\OLL\TXT\comedies\All's Well that Ends Well.txt renamed as Bacon0011.txt
C:\temp\shakes\OLL\TXT\comedies\As You Like It.txt renamed as Bacon0012.txt
C:\temp\shakes\OLL\TXT\comedies\Cymbeline.txt renamed as Bacon0013.txt
C:\temp\shakes\OLL\TXT\comedies\Love's Labour's Lost.txt renamed as Bacon0014.txt
C:\temp\shakes\OLL\TXT\comedies\Measure for Measure.txt renamed as Bacon0015.txt
C:\temp\shakes\OLL\TXT\comedies\Much Ado About Nothing.txt renamed as Bacon0016.txt
C:\temp\shakes\OLL\TXT\comedies\Pericles, Prince of Tyre.txt renamed as Bacon0017.txt
C:\temp\shakes\OLL\TXT\comedies\The Comedy of Errors.txt renamed as Bacon0018.txt
```

so after pressing *Rename* the texts have been renamed **Bacon** plus an incrementing number formatted to 4 digits.

See also : compare two files 347, file chunker 348, find duplicates 348

10.6.8 move files to sub-folders

This function allows you to take a whole set of files in a folder and move them to suitable sub-folders.

Example:

In **c:\temp** you have

```
2001 Jan.txt
2001 Feb.txt
2003 Jan.txt
2003 Feb.txt
2003 March.txt
2003 Oct.txt
```

etc. and you want them sorted by year into different folders.

Using the template **AAAA*** you will take the first four characters of your files and place each into a sub-folder named appropriately.

Results

c:\temp\2001 contains **2001 Jan.txt, 2001 Feb.txt**

and all the others are in **`c:\temp\2003`**

Syntax

? = ignore this character

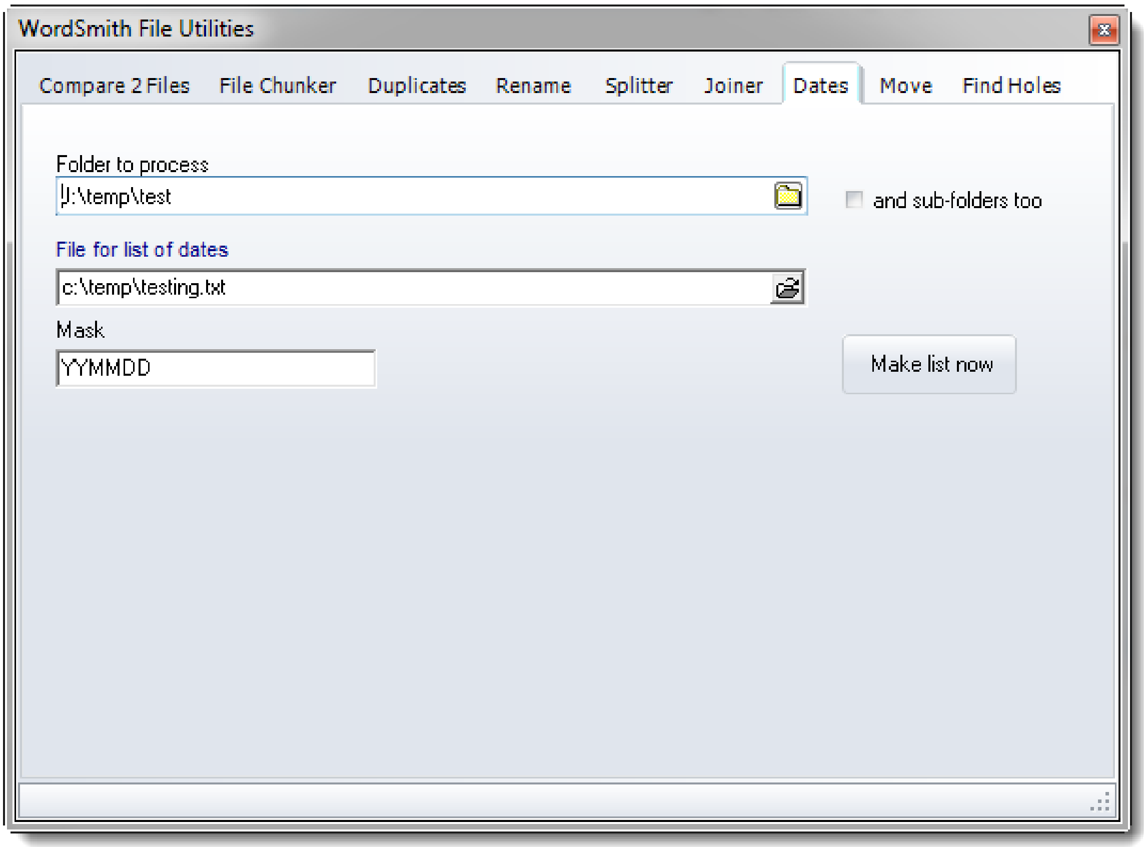
A = use this character in the file-name

* = use no further characters in the file-name

10.6.9 dates and times

Purpose

The aim here is to parse your file-names identifying suitable textual file dates and times 126, where you have incorporated something suitable in the file-name. Suitable dates can be re-used by saving file-choices as favourites 50.



Folder to process: J:\temp\test

□ and sub-folders too

File for list of dates: c:\temp\testing.txt

Mask: YYMMDD

Mask Syntax

The procedure reads any file-names in the *Folder to process* (and optionally its sub-folders) and attempts to parse them. If an indicator is found it will record a suitable date combination.

Suitable indicators of textual date are

YY or YYYY	year, two or four digits (YY=a 20th Century date)

MM	month
DD	day
*	skip all characters until a digit is found

The procedure doesn't understand words such as "December" or "Five", it only uses digits. Any character other than Y,M,D,* in the mask simply gets ignored.

Output

The program will always add each entry found to a simple text file (*File for list of dates*) listing its file-name and adding a suitable date as expected in the auto-date procedure 48, (or **`<no date found>`** if the mask didn't match a valid date). In addition, where the result is 1st January 1980 or later, it will set the file's time and date in the operating system to the date as parsed, so that WordSmith will automatically match the date of the text contents to the date stored on disk.

When all files have been processed, the program opens the list of files in Notepad or equivalent. Use it afterwards in the auto-date procedure 48 within file-choosing and save your preferred text files as favourites 50.

Examples

Your Mask	Source file	Date and Time interpreted
`YYYYMMDD`	**`20060512 Peter monologue.txt`**	12th May 2006 (first 8 characters used in the mask)
`YYMMDD`	**`841231.txt`**	31st December 1984 (20th Century assumed if YY mask used)
`DDMMYY`	**`311284.txt`**	31st December 1984
`DDMMYYYY`	**`20060512 Peter monologue.txt`**	20th June, the year 512 AD
`DDMM`	**`20060512 Peter monologue.txt`**	20th June of the current year
`######YYYYMMDD`	**`Peter 20060512.txt`**	12th May 2006 (first six characters were ignored, five for Peter, one for space)
`*YYYY`	**`Peter 20060512.txt`**	15 July 2006 (all characters to first digit skipped, then next 4 used for year date)
`YYYY`	**`1086 Domesday book.txt`**	15 July 1086 (there were only four digits)
`YYYYMMDD`	**`1086 Domesday book.txt`**	15 July 1086 (mask had 8 digits but file-name only 4)
`YYYY#MM#DD`	**`2006,05/12,10-54.txt`**	12th May 2006
`YYYY MM DD`	**`2006,05/12,10-54.txt`**	12th May 2006

10.6.10 find holes in texts

After text files have been copied from one source to another, they may get slightly corrupted with holes 466 in the stream of text. This utility lets you seek out the texts in your corpus which have got corrupted in this way and optionally lets you delete them. If you want to convert the holes to space-characters, use the Text Converter.

10.7 Text Converter

10.7.1 purpose

This program does a "Search & Replace", on virtually any number of files.

It is very useful for going through large numbers of texts and re-formatting them as you prefer, e.g. taking out unnecessary spaces, ensuring only paragraphs have <Enter> at their ends, changing accented characters, ensuring you have Windows **£** symbols, etc.

converting text

For a simple search-and-replace you can type in the search item and a replacement; for more complex conversions, use a Conversion File 362 so that **Text Converter** knows which symbols or strings to convert. It operates under Windows and saves using the Windows character set 419, but will convert text using DOS or Windows character sets. You can use it to make your text files suitable for use with your Internet browser.

It does a "search and replace" much as in word-processors, but it can do this on lots of text files, one after the other. As it does so, it can also replace up to **any number of** strings, not just one.

Once the conversion file is prepared and Settings 355 specified, the **Text Converter** will read each source file and either create a new version or replace the old one, depending on the over-write setting 355.

You will be able to see the details of how many instances of each string were found and replaced overall.

filtering files

And/or you may need to make sure texts which meet certain criteria are put into the right folders 360.

Tip

The easiest way to ensure your text files are the way you want, especially if you have a very large number to convert, is to copy a few into a temporary folder and try out your conversion file with the Text Converter. You may find you've failed to specify some necessary conversions. Once you're sure everything is the way you want it, delete the temporary files.

See also: Text Converter Contents 355, The buttons 441

10.7.2 index

Explanations

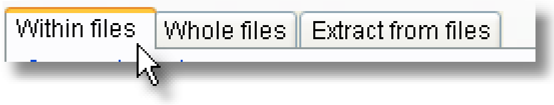
What is the Text Converter and what's it for? 354
Getting Started... 355
Convert the text format 365
Filters 360
Sample Conversion File 364
Syntax 363
Conversion File 362

See also : WordSmith Main Index 2

10.7.3 settings

1. Choose *Files* (the top left tab). Decide whether you want the program to process sub-folders of the one you choose. There is no limit to the number of files Text Converter can process in one operation.
2. Click on the *Conversion* or *Filters* 360 tab, and:
3. Decide whether you want to make copies of the text files, or to over-write the originals. Obviously you must be confident of the changes to choose to over-write; copying however may mean a problem of storage space.

Choose between "Within files", "Whole files" or "Extract from files"



Within files = make some alterations to specific words in each text file, if found
For example, specify what to convert, that is the search-words and what you want them to be replaced with. For a quick conversion you can simply type in a word you want to change and its replacement (e.g. *Just one change* so that **`responsable`** becomes **`responsible`**) or you can choose your own pre-prepared Conversion File 362.

Whole files = make some alterations affecting all the words in each text file
E.g. in the Whole Files section you can choose simply to update legacy files 365 in various ways, e.g. by choosing

Dos to Windows,
Unix to Windows,
MS Word .doc to .txt,
into Unicode,
etc.

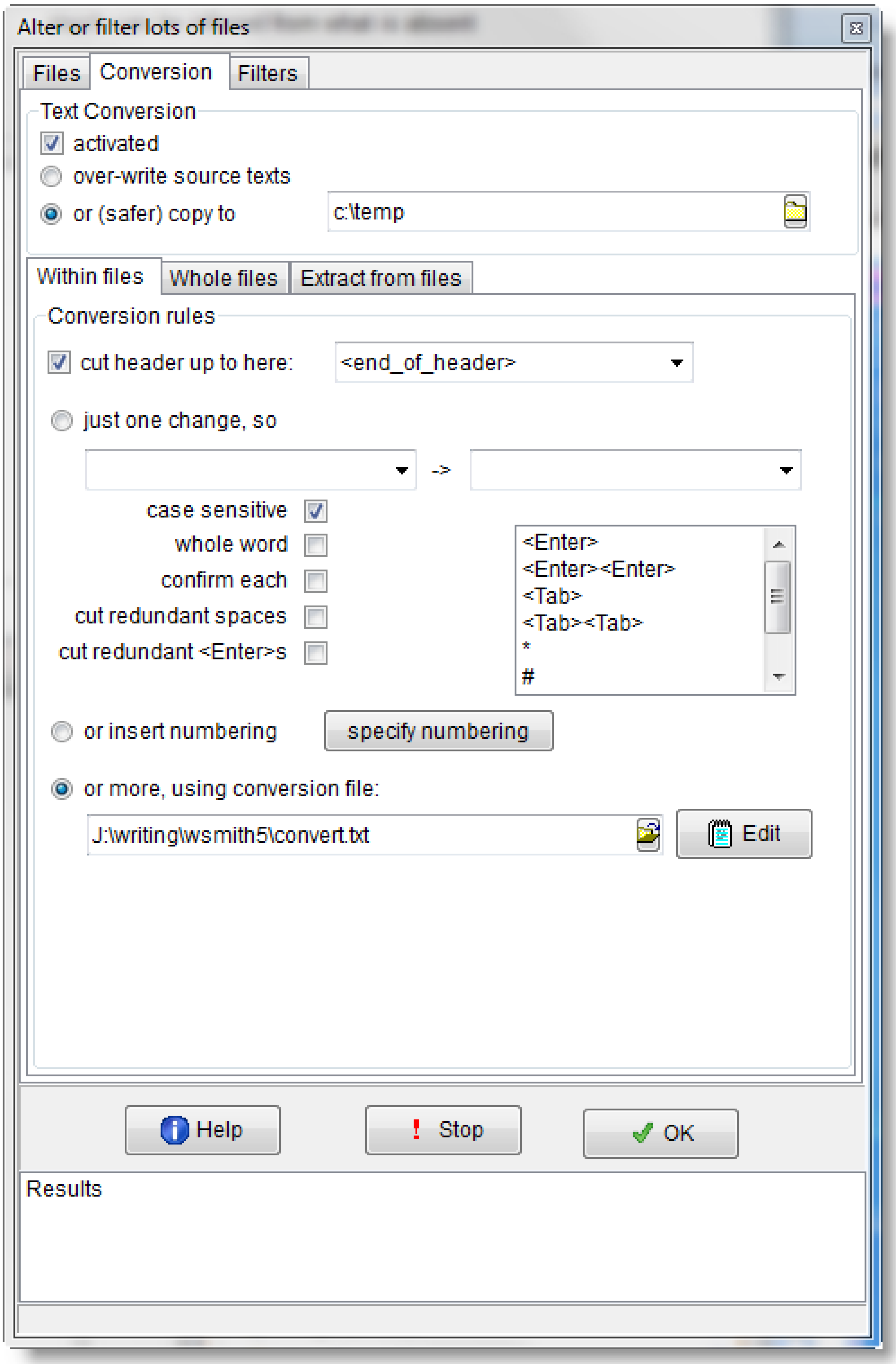
Or if you want simply to extract some text from your files, you should choose the Extract from files

359 tab.

If you might want some files not to be converted, or simply don't want any conversions but instead to place files in appropriate sub-folders, choose the Filters 360 tab at the top.

If you choose *Over-write Source texts,* Text Converter will work more quickly and use less disk space, but of course you should be quite sure your conversion file codes are right before starting! See copy to 358 for details of how the folders get replicated in a copy operation.

Note that ***some space on your hard disk will be used even if you plan to over-write***. The conversion process does its work, then if all is well the original file is deleted, and the new version copied. There has to be enough room in the destination folder for the largest of your new files; it is much quicker for it to be on the same drive as the source texts. If it isn't, your permission will be asked to use the same drive.



Alter or filter lots of files

Text Conversion

☑ activated

○ over-write source texts

● or (safer) copy to c:\temp

Conversion rules

☑ cut header up to here: <end_of_header>

○ just one change, so ->

case sensitive ☑

whole word □

confirm each □

cut redundant spaces □

cut redundant <Enter>s □

<Enter>
<Enter><Enter>
<Tab>
<Tab><Tab>
*
#

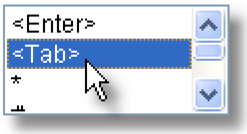
○ or insert numbering

● or more, using conversion file:

J:\writing\wsmith5\convert.txt

Results

inserting <Tab>, <Enter> etc

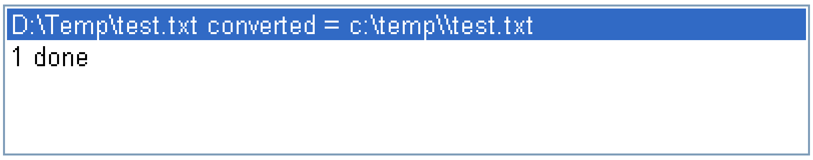


Choose in the listbox and drag to one of the windows to left or right of ->. The string inserted will conform to the format 362.

cutting out a header from each file

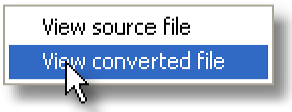
It can be useful to get a header removed. In the screenshot example, any text which contains **</teiHeader>** will get all the beginning of the file up to that point cut out.

Press *OK* to start; you will see a list of results, as in the screenshot below.
If you want to stop **Text Converter** at any time, click on the Stop button or press Escape.



```
D:\Temp\test.txt converted = c:\temp\\test.txt
1 done
```

Right-click to see the source or the converted result file:



See also: Text Converter Contents 355.

10.7.3.1 Text Converter: copy to

If you choose to copy the files you are converting, instead of converting or filtering them in place, which is a lot safer, the new files created will be structured like this.

Suppose you are processing **d:\texts\2007\literature** and copying to **c:\temp** and suppose **d:\texts\2007\literature** contains this sort of thing:

```
d:\texts\2007\literature\shakespeare\hamlet.pdf
d:\texts\2007\literature\shakespeare\macbeth.pdf
...
d:\texts\2007\literature\shakespeare\poetry\sonnet1.pdf
d:\texts\2007\literature\shakespeare\poetry\sonnet2.pdf
...
d:\texts\2007\literature\french\victor hugo\miserables.pdf
d:\texts\2007\literature\french\poetry\baudelaire\le chat.pdf
...
```

you will get

```
c:\temp\shakespeare\hamlet.txt
c:\temp\shakespeare\macbeth.txt
...
```

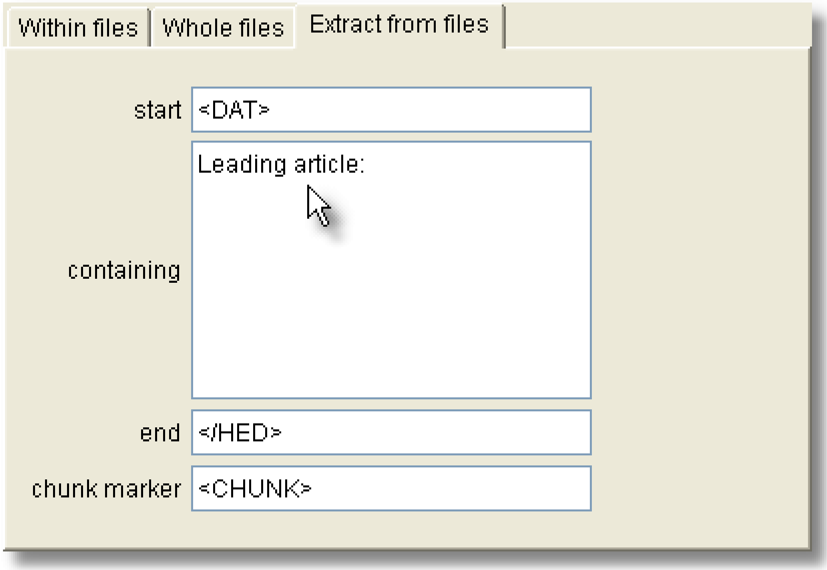
```
c:\temp\shakespeare\poetry\sonnet1.txt
c:\temp\shakespeare\poetry\sonnet2.txt
...
c:\temp\french\victor hugo\miserables.txt
c:\temp\french\poetry\baudelaire\le chat.txt
...
```

In other words, for each file successfully converted or filtered, any same directory structure beyond the starting point (**`d:\texts\2007\literature`** in the example above) will get appended to the destination.

10.7.4 extracting from files

The point of it...

The idea is to be able to extract something useful from within larger files. In the example below, I wanted to extract the headlines only from some newspaper text. I knew that the header for each text contained **`<DAT>`** (date of publication mark-up) and that the headline ended **`</HED>`**, and I wanted only those chunks which contained the phrase **`Leading article:`**.



start <DAT>

containing Leading article:

end </HED>

chunk marker <CHUNK>

The results I got looked like this:

<CHUNK "1"><DAT>05 August 2001</DAT>
<SOU>The Observer</SOU>
<PAG>26</PAG>
<HED>Comment: Leading article: Ealing's lessons: Time for steel from the peacemakers</HED></CHUNK>
<CHUNK "2"><DAT>05 August 2001</DAT>
<SOU>The Observer</SOU>
<PAG>26</PAG>
<HED>Comment: Leading article: The free market can't house us all: Why Government has to intervene</HED></CHUNK>

<CHUNK "3"><DAT>05 August 2001</DAT>
<SOU>The Observer</SOU>
<PAG>26</PAG>
<HED>Comment: Leading article: What a turn-on: Cat's whiskers are the bee's knees</HED></CHUNK>

Settings

containing : **all** non-blank lines in this box will be required. Leave it blank if you have no requirement that the chunk you want to extract contains any given word or phrase.
chunk marker : Leave blank, otherwise each chunk will be marked up as in the example above, if it begins with **<** and ends with **>**. The reason for this marker is to enable subsequent splitting 343.

10.7.5 filtering: move if

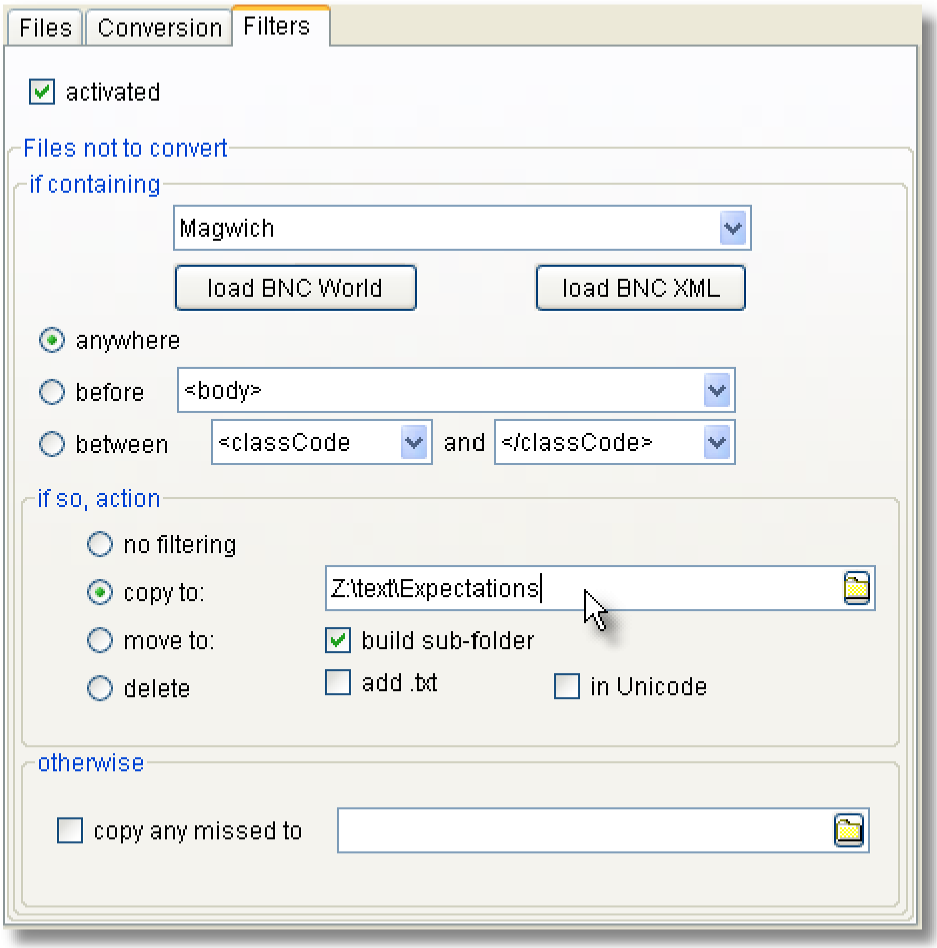
This function allows you to specify a word or phrase, look for it in each file, and if it's found move that file into a new folder.

The point of it ...

Suppose you have a whole set of files some of which contain dialogues between Pip and Magwich, others containing references to the Great Wall of China or the anatomy of fleas. You want those with the Pip-Magwich dialogues and you want them to go into a folder called *Expectations*.

How to do it

1. Click on the *Filters* tab (at the top).
2. Now the *Activated* checkbox.



☑ activated

Files not to convert

if containing

Magwich

● anywhere

○ before <body>

○ between <classCode and </classCode>

if so, action

○ no filtering

● copy to: Z:\text\Expectations

○ move to:

☑ build sub-folder

○ delete

□ add .txt

□ in Unicode

otherwise

□ copy any missed to

3. Specify a word or phrase the text must contain. This is case sensitive. In this case **`Magwich`** has been specified.

4. Choose whether that word or phrase has to be found

- anywhere in the text,
- anywhere before some other word or phrase, or
- between 2 different words or phrases.

5. Decide what happens if the conditions are met:

- nothing, i.e. ignore that text file
- copy to a certain folder, or
- move to that folder, or
- delete the file (careful!).

Action options

- You can also decide to build sub-folder(s) based on the word(s) or phrase(s) you chose in #3. (The idea is to get your corpus split up useful sub-folders whose names mean something to you.) If *build sub-folder* is not checked everything goes into the *copy to* or *move to* folder.
- And you may have the program add **`.txt`** (useful if as with the BNC World Edition there are no file extensions) and/or convert it to Unicode.
- You could also have any texts not containing the word **`Magwich`** copied to a specified folder.

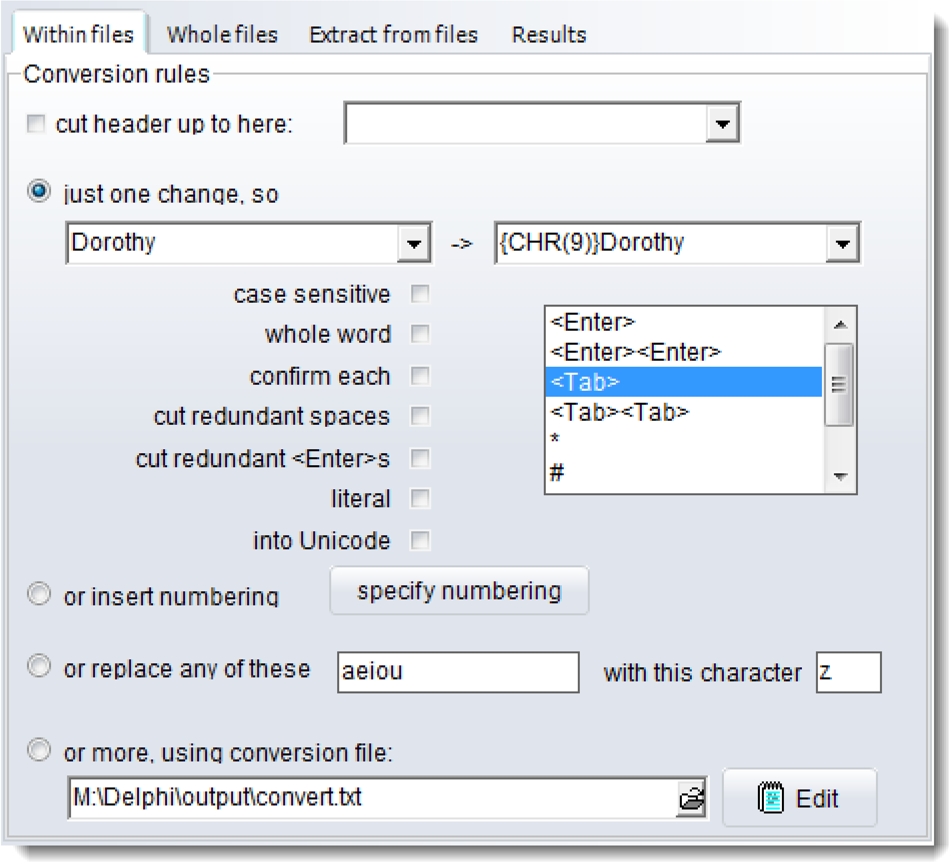
The *load BNC World* and *load BNC XML* buttons are specific to those two editions of the BNC and read text files with similar names which you will find in your **`Documents\wsmith6`** folder.

See also: Text Converter Contents 355.

10.7.6 Convert within the text file

Your choices here are 5:

1. cut out a header



Conversion rules

☐ cut header up to here:

● just one change, so

Dorothy -> {CHR(9)}Dorothy

case sensitive ☐
whole word ☐
confirm each ☐
cut redundant spaces ☐
cut redundant <Enter>s ☐
literal ☐
into Unicode ☐

<Enter>
<Enter><Enter>
<Tab>
<Tab><Tab>
*
#

○ or insert numbering

○ or replace any of these aeiou with this character z

○ or more, using conversion file:

M:\Delphi\output\convert.txt

and/or

2. make one change only
3. insert numbering
4. replace some problem characters
5. use a script to determine a whole set of changes. There is an sample 364 to see.

10.7.6.1 conversion file

Prepare your Text Converter conversion file using a plain text editor such as Notepad. You could use **`Documents\wsmith6\convert.txt`** as a basis.

If you have accented characters 419 in your original files, use the DOS editor to prepare the conversion file if they were originally written under DOS and a Windows editor if they were written in a Windows word-processor. Some Windows word processors can handle either format.

There can be any number of lines for conversion, and each one can contain two strings, delimited with " " quotes, each of up to 80 characters in length.
The Text Converter makes all changes in order, as specified in the Conversion File. Remember one alteration may well affect subsequent ones.

Alterations that increase the original file

Most changes reduce the size of an original. But Text Converter will cope even if you need to increase the original file -- as long as there's disk space!

Tip

To get rid of the <Enter> at line ends but not at paragraph ends, first examine your paragraph ends to see what is unique about them. If for example, paragraphs end with two <Enters>, use the following lines in your conversion file:
`"{CHR(13)}{CHR(10)}{CHR(13)}{CHR(10)}" -> "{%%}"`
(this line replaces the two <Enters> with **`{%%}`** .) (It could be any other unique combination. It'll be slightly faster if you make the search and the replacement the same length, as in this case, 4 characters)
`"{CHR(13)}{CHR(10)}" -> " "`
(this line replaces all other <Enters> with a space, to keep words separate)
`"{%%}" -> "{CHR(13)}{CHR(10)}{CHR(13)}{CHR(10)}"`
(this line replaces the **`{%%}`** combination with <Enter><Enter>, thus restoring the original paragraph structure)
`/S`
(this line cuts out all redundant spaces)

See also: sample conversion file 364, syntax 363, Text Converter Contents 355.

10.7.6.2 syntax

The syntax for a Conversion File 362 is:

- Only lines beginning / or " are used. Others are ignored completely.
- Every string for conversion is of the form "A" -> "B". That is, the original string, the one you're searching for, enclosed in double quotes, is followed by a space, a hyphen, the > symbol, and the replacement string.
- You can use " (double quotes) and hyphen where you like without any need to substitute them, but for obvious reasons there must not be a sequence like **`" -> "`** in your search or replace string.

Removing all tags

To remove all tags, choose **`"<*>" -> ""`** as your search string.

Control Codes

Control codes can be symbolised like this: {CHR(xxx)} where xxx is the number of the code. Examples: **`{CHR(13)}`** is a carriage-return, **`{CHR(10)}`** is a line-feed, **`{CHR(9)}`** is a tab. To represent <*Enter*> which comes at the end of paragraphs and sometimes at the end of each line, you'd type **`{CHR(13)}{CHR(10)}`** which is carriage-return followed immediately by line-feed.

Use **{CHR(34)}** if you need to refer to double inverted commas. See search-word syntax 159 for more.

Wildcards

The search uses the same mechanism that Concord uses. You may use the same wildcards as in Concord search-word syntax 159. By default the search-and-replace operates on whole words.

Examples:

"book" -> "bk" will replace **book** with **bk** but won't replace **books** or **textbook**

"*book" -> "bk" will replace **book** or **textbook** with **bk** but won't replace **books** or **textbooks**

"book*" -> "bk" will replace **book** or **books** with **bk** but won't replace **textbook** or **textbooks**

To show a character is to be taken literally, put it in quotes (e.g. "*","<"). See below for use of the / L parameter.

Unbounded, case Insensitive, Confirm, redundant Spaces, redundant <Enter>s

/C stops to confirm you wish to go ahead before each change.

/U does an unbounded search (ensuring the alteration happens whether there's a word separator 427 on either side or not) (/U "the" finds *the* but also finds *other, then* and *bathe*).

/I does a case insensitive search (/I "restaurant" -> "hotel" replaces *restaurant* with *hotel* and *RESTAURANT* with *HOTEL* and *Restaurant* with *Hotel*, i.e. respecting case as far as possible).

You can combine these, e.g.

```
 /IC "the" -> "this"
```

/S cuts out all redundant spaces. That is, it will reduce any sequence of two or more spaces to one, and it also removes some common formatting problems such as a lone space after a carriage-return or before punctuation marks such as .,; and). **/S** can be used on a line of its own or in combination with other searches.

/E cuts out all redundant <Enter>s. That is, it will reduce any sequence of two or more carriage-return+line-feeds (what you get when you press Enter or Return) to one. **/E** can be used on a line of its own or in combination with other searches.

/L means both the search and replace strings are to be taken as literal. (Normally a sequence like **<#*>** would need quotes around each character because **< >** are mark-up signals and # and ***** are special wildcard characters, thus **"<""#""*"">"** which is tricky! Put **/L** at the start of the line to avoid this.)

See **Documents\wsmith6 \convert.txt** to see examples in use.

See also: Text Converter Contents 355.

10.7.6.3 sample conversion file

You could copy all or part of this to the clipboard 422 and paste it into notepad.

```
[ comment line -- put whatever you like here, it'll be ignored ]

[ first a spelling correction ]
```

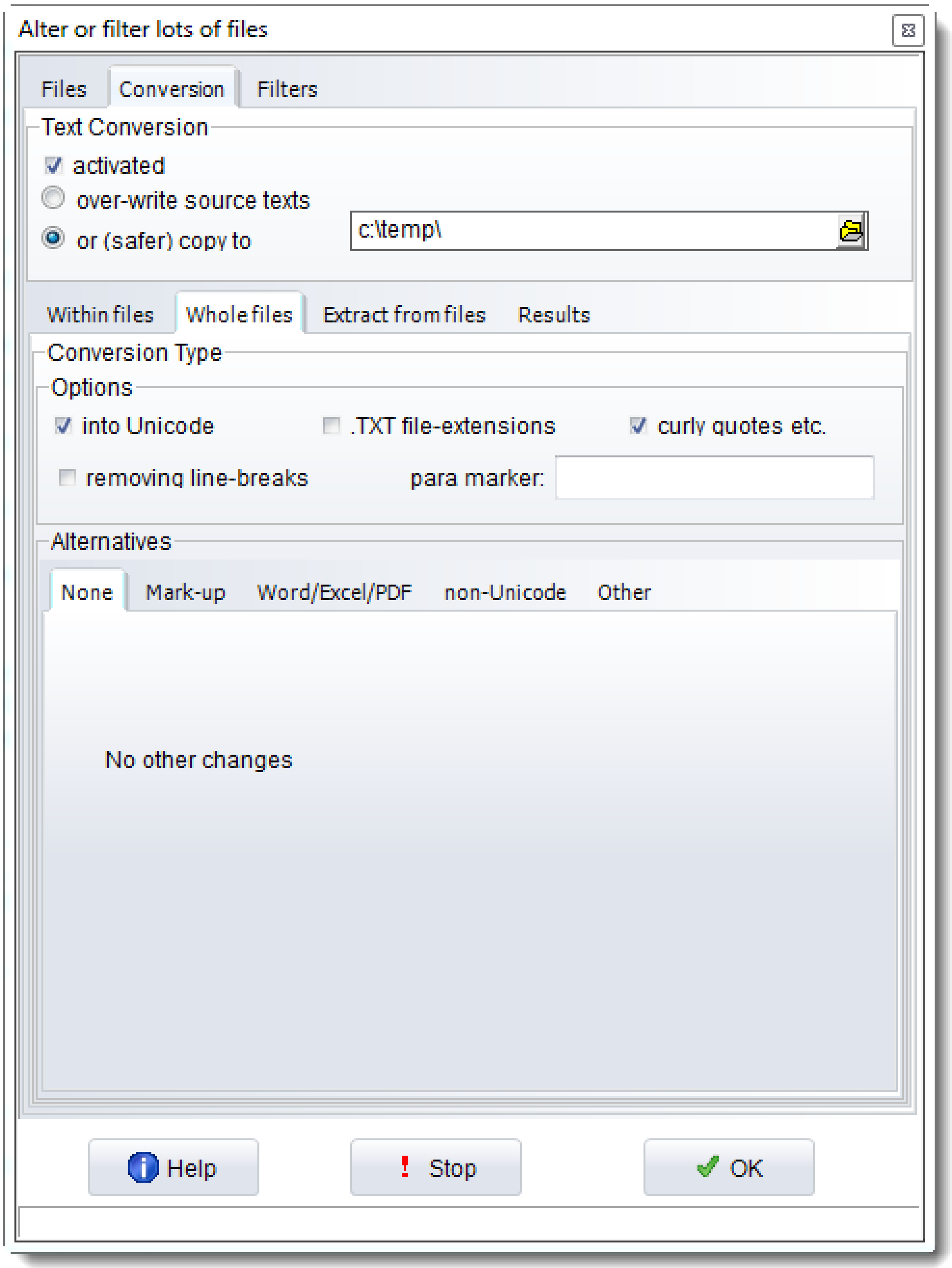
```
"responsable" -> "responsible"

  [ now let's change brackets from < > to [ ] and { } to ( ) ]
"*<*" -> "["
"*>*" -> "]"
"*}*" -> ")"
"*{*" -> ")"
/S
  [ that will clear all redundant spaces]
```

The file **`Documents\wsmith6\convert.txt`** is a sample conversion file for use with British National Corpus text files.

See also: Text Converter Contents 355.

10.7.7 Convert format of entire text files

To convert a series of whole text files from one format to another, choose one or more of these options:



Alter or filter lots of files

Files | Conversion | Filters

Text Conversion

☑ activated

○ over-write source texts

● or (safer) copy to c:\temp\

Within files | Whole files | Extract from files | Results

Conversion Type

Options

☑ into Unicode □ .TXT file-extensions ☑ curly quotes etc.

□ removing line-breaks para marker:

Alternatives

None | Mark-up | Word/Excel/PDF | non-Unicode | Other

No other changes

These formats allow you to convert into formats which will be suited to text processing.

into Unicode:

.... this is a better standard than ASCII or ANSI as it allows many more characters to be used, suiting lots of languages. See Flavours of Unicode 430.

TXT file extensions:

... makes the filename end in **.txt** (so that Notepad will open without hassling you; Windows was baffled by the empty filenames of the BNC editions prior to the XML edition). If you choose this you will be asked whether to force **.txt** onto all files regardless, or only ones which have no file extension at all.

curly quotes etc.:

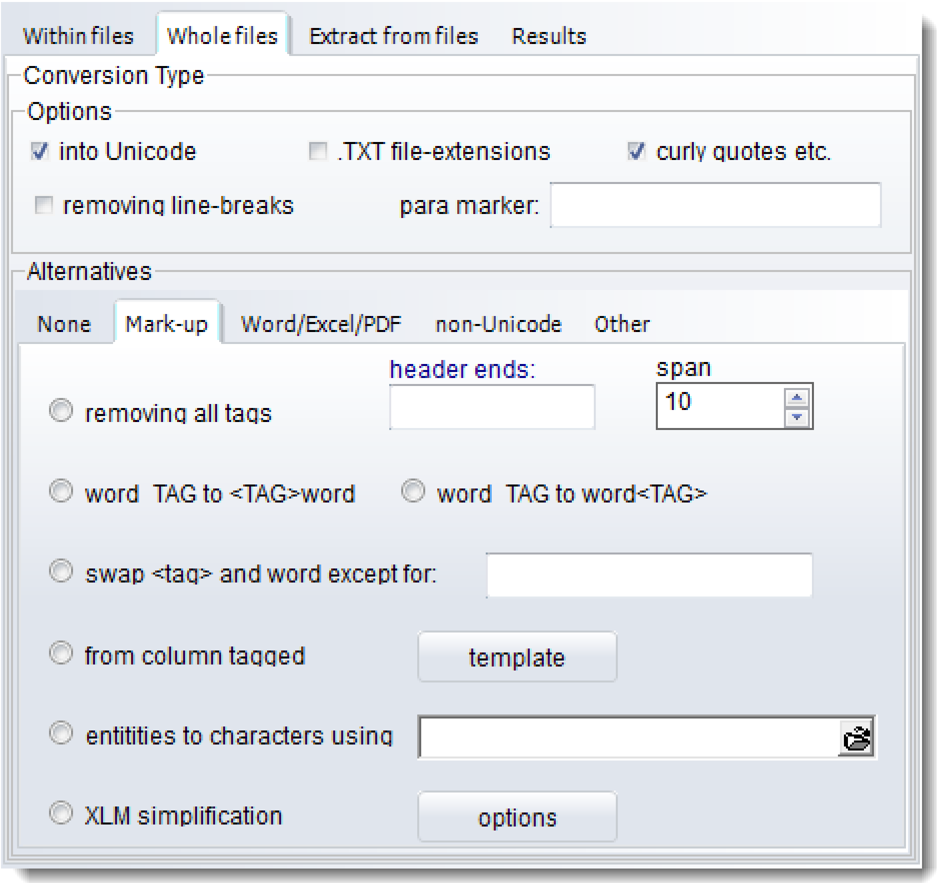
... changes any curly single or double quote marks or apostrophes into straight ones, ellipses into three dots, and dashes into hyphens. (Microsoft's curly apostrophes differ from straight ones.)

removing line-breaks

... replaces every end of line line-break with a space. Preserves any true paragraph breaks, which you must ensure are defined (default = **<Enter><Enter>** -- in other words two line-breaks one after the other with no words between them).

See also: Mark-up 368, Word/Excel/PDF 371, non-Unicode 372, convert within text files 362, MS Word documents 444, Guide to handling the BNC

10.7.7.1 Mark-up changes



☑ into Unicode □ .TXT file-extensions ☑ curly quotes etc.

□ removing line-breaks para marker:

○ removing all tags header ends: span 10

○ word TAG to <TAG>word ○ word TAG to word<TAG>

○ swap <tag> and word except for:

○ from column tagged

○ entitities to characters using

○ XLM simplification

removing all tags

would convert **The<DT><the> TreeTagger<NP><TreeTagger> is<VBZ>**... into **The Treetagger is**. Can plough through a copy of the whole BNC, for example, and make it readable. If you have specified a header string it will cut the header up to that point too. Uses the selected span for looking for the next **>** when it finds a **<**.

word_TAG to <TAG>word

The Helsinki corpus can come tagged like this (COCOA tags)

```
   the_D occasion_N of_P her_PRO$ father's_N$ death_N
```

and this conversion procedure will change it to

```
   <D>the <N>occasion <P>of <PRO$>her <N$>father's <N>death
```

Note: this procedure does not affect underscores within existing <> markup.

word_TAG to word<TAG>

converts text like

```
It_PP is_VBZ easy_JJ
```

or Stanford Log-linear POS tagger output like

```
It/PP is/VBZ easy/JJ
```

to

```
It<PP> is<VBZ> easy<JJ>
```

You will have to confirm which character such as _ or / divides the word from the tags. Note: before it starts, it will clear out any existing <> markup.

swap tag and word

converts text like

```
It<PP> is<VBZ> easy<JJ>
```

to

```
<PP>It <VBZ>is <JJ>easy
```

or vice-versa. In other words swapping the order of tags and words. The procedure effects a swap at each space in the non-tagged text sequence.

Any tags which do not qualify a neighbouring word but for example a whole sentence or a paragraph should not be swapped, so fill in the box to the right with any such tags, using commas to separate them, e.g. **`<s>,</s>,<p>,</p>`**

from column tagged

The Stuttgart Tree Tagger produces output like this separating 3 aspects of each word with a <tab>:

```
word          pos      lemma
The           DT       the
TreeTagger    NP       TreeTagger
is            VBZ      be
easy          JJ       easy
to            TO       to
use           VB       use
.             SENT     .
```

You will need to supply a template for your conversion.

Template syntax and examples:

1. Any number in the template refers to the data in that column number. (**`The`** is in column 1 above, **`DT`** in column 2 of the original.)
2. Only columns mentioned in the template get used in the final output.
3. Separate columns in your template with a / slash.
4. You can add letters and symbols if you like.
5. A space will get added after each line of your original.

Examples:

- the template **`1/<3>/<2>`** will produce with the cases above **`The<the><DT>`**

`Treetagger<Treetagger><NP> is<be><VBZ>` etc.

- the template **`<POS="2">/1`** will produce **`<POS="DT">The <POS="NP">Treetagger <POS="VBZ">is`** etc.

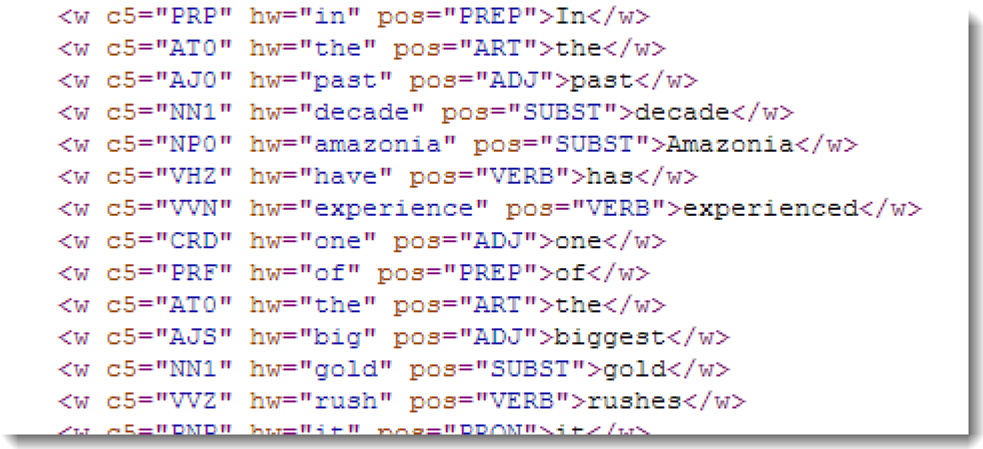
It will present the text as running text, no longer in columns, but with a break every 80 characters.

entities to characters

... converts HTML or XML symbols which are hard to read such as **`&eacute;`** to ones like **é**. Specify these in a text file. There is a sample file pre-prepared for you, **`html_entities.txt,`** in your Documents\wsmith6 folder; look inside and you'll see the syntax.

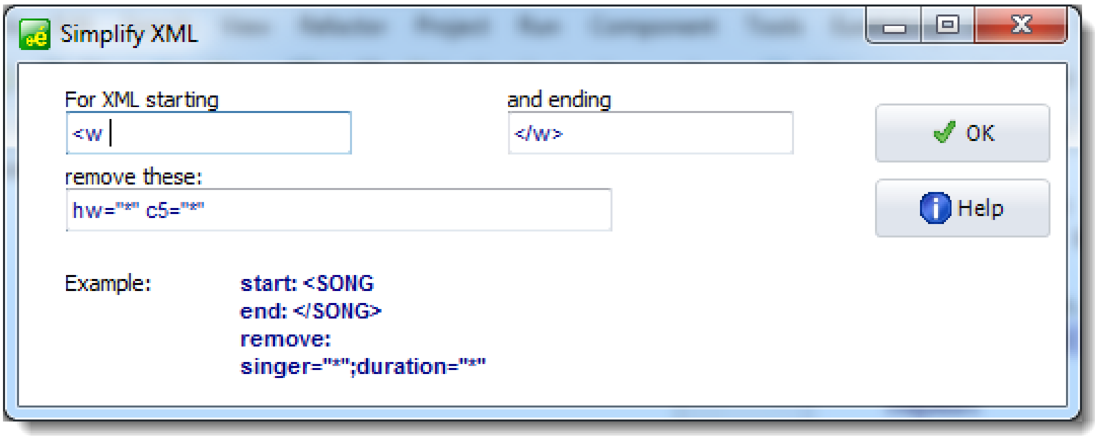
XML simplification

The idea is to remove any mark-up in XML data which you really do not wish to keep. For example, in the BNC XML edition you might wish to keep only the **`pos="*"`** mark-up and remove the **`c5`** and **`hw`** attributes.



```
<w c5="PRP" hw="in" pos="PREP">In</w>
<w c5="AT0" hw="the" pos="ART">the</w>
<w c5="AJ0" hw="past" pos="ADJ">past</w>
<w c5="NN1" hw="decade" pos="SUBST">decade</w>
<w c5="NP0" hw="amazonia" pos="SUBST">Amazonia</w>
<w c5="VHZ" hw="have" pos="VERB">has</w>
<w c5="VVN" hw="experience" pos="VERB">experienced</w>
<w c5="CRD" hw="one" pos="ADJ">one</w>
<w c5="PRF" hw="of" pos="PREP">of</w>
<w c5="AT0" hw="the" pos="ART">the</w>
<w c5="AJS" hw="big" pos="ADJ">biggest</w>
<w c5="NN1" hw="gold" pos="SUBST">gold</w>
<w c5="VVZ" hw="rush" pos="VERB">rushes</w>
<w c5="PNP" hw="it" pos="PRON">it</w>
```

To do so, press the Options button and complete for example like this:



Simplify XML

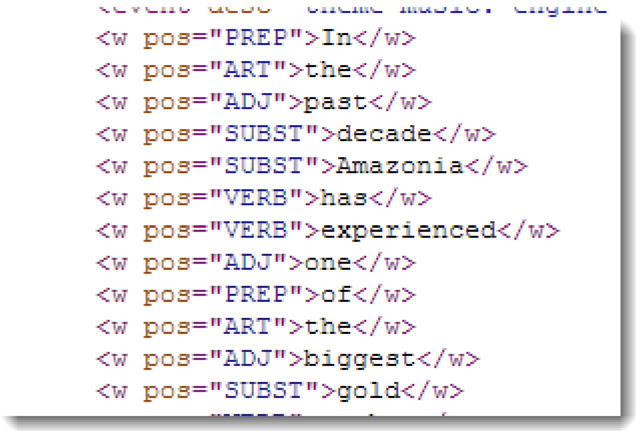
For XML starting <w

and ending </w>

remove these: hw="*" c5="*"

Example:
start: <SONG
end: </SONG>
remove:
singer="*";duration="*"

resulting in a saved XML file with a structure like this:

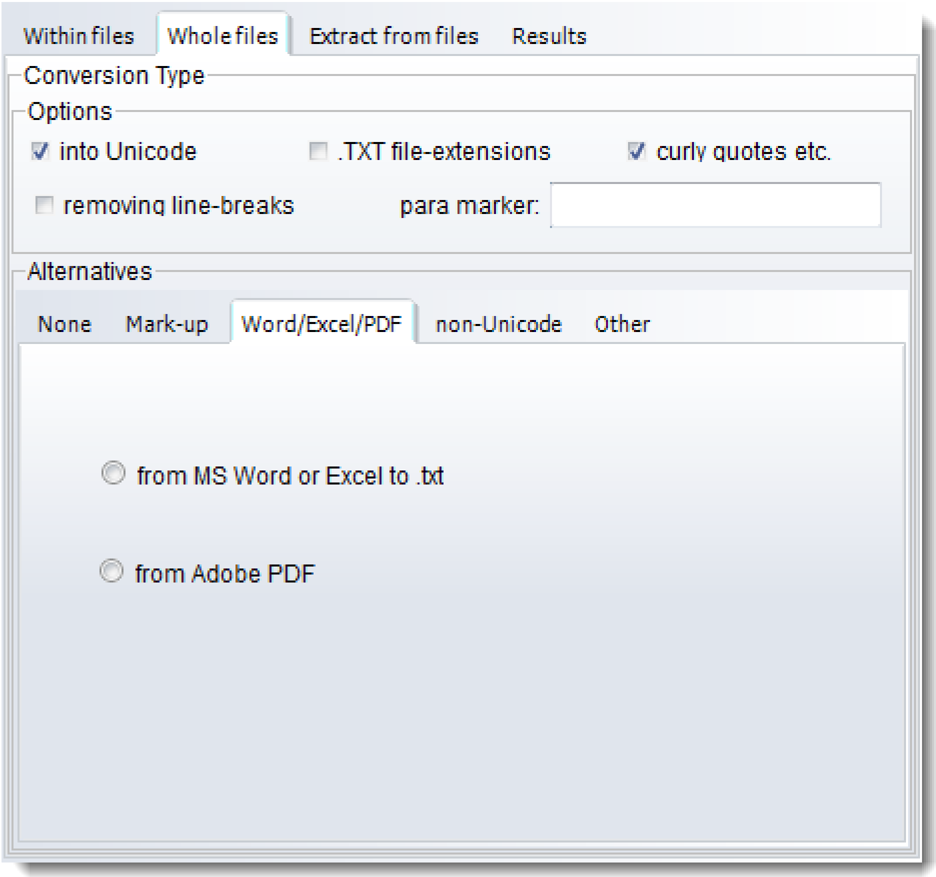


```
<w pos="PREP">In</w>
<w pos="ART">the</w>
<w pos="ADJ">past</w>
<w pos="SUBST">decade</w>
<w pos="SUBST">Amazonia</w>
<w pos="VERB">has</w>
<w pos="VERB">experienced</w>
<w pos="ADJ">one</w>
<w pos="PREP">of</w>
<w pos="ART">the</w>
<w pos="ADJ">biggest</w>
<w pos="SUBST">gold</w>
```

The procedure simply looks for all sections which begin and end with the required strings and delete any sections in between which contain the strings you specify in the *remove these* section. No further account of context is taken. Note that the order of attributes is not important, so we could have specified **c5="*"** first.

See also: Convert Entire Texts 365

10.7.7.2 Word, Excel, PDF



Within files | Whole files | Extract from files | Results

Conversion Type

Options

☑ into Unicode □ .TXT file-extensions ☑ curly quotes etc.

□ removing line-breaks para marker:

Alternatives

None | Mark-up | Word/Excel/PDF | non-Unicode | Other

○ from MS Word or Excel to .txt

○ from Adobe PDF

from MS Word or Excel to .txt

This is like using "Save as Text" in Word or Excel. Handles .doc, .docx 444 (Office 2007) and .xls files.

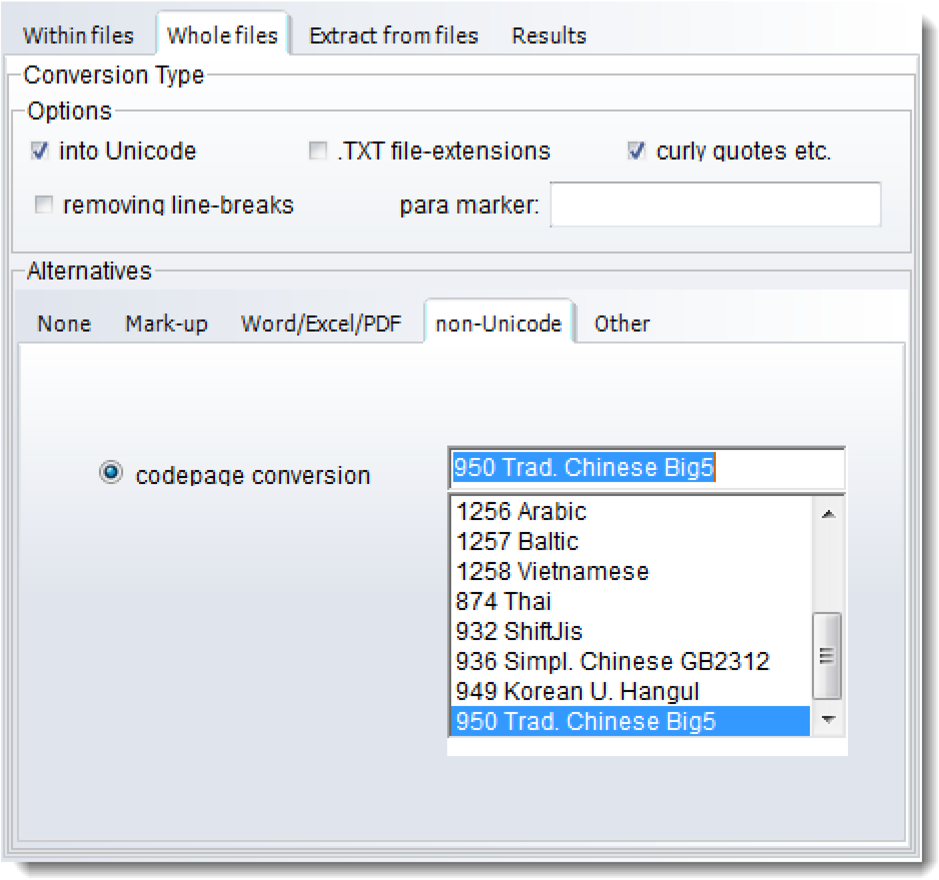
from PDF

... into plain text. Not guaranteed to work with every .PDF as formats have changed and some are complex.

To convert PDFs to plain text can be extremely tricky even if you own a licensed copy of the Adobe software (Adobe themselves created the PDF format in 1993). That is because PDF is a representation of all the dots, colours, shapes and lines in a document, not a long string of words. It can be very hard with an image of the text, to determine the underlying words and sentences. A second problem is that PDFs can be set with security rights preventing any copying, printing, editing etc. Other formats (.TXT, .DOC, .DOCX, .XML, .HTML, .RTF etc.) are OK in principle as they do not contain only an image but also store within themselves the words and sentences.

See also: Convert Entire Texts 365

10.7.7.3 non-Unicode Text



Conversion Type

Options

☑ into Unicode ☐ .TXT file-extensions ☑ curly quotes etc.

☐ removing line-breaks para marker:

● codepage conversion: 950 Trad. Chinese Big5

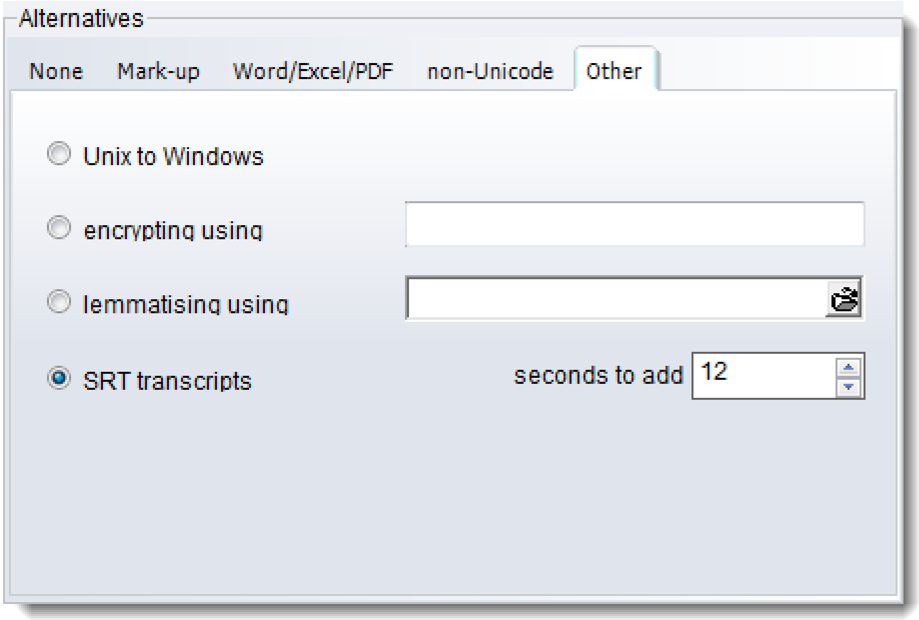
1256 Arabic
1257 Baltic
1258 Vietnamese
874 Thai
932 ShiftJis
936 Simpl. Chinese GB2312
949 Korean U. Hangul
950 Trad. Chinese Big5

Codepage conversion

This allows you to convert 1-byte based formats, for example from Chinese Big5 or GB2312, Japanese ShiftJis, Korean Hangul to Unicode.

See also: Convert Entire Texts 365

10.7.7.4 Other changes



○ Unix to Windows

○ encrypting using

○ lemmatising using

● SRT transcripts   seconds to add 12

Unix to Windows

Unix-saved texts don't use the same codes for end-of-paragraph as Windows-saved ones.

encrypting using

... allows you to encrypt your text files. You supply your own password in the box to the right. When WordSmith processes your text files, e.g. when running a concordance it will restore the text as needed but otherwise the text will be unintelligible. Encrypted files get the file extension **`.WSencrypted`**. For example, if your original is **`wonderful.txt`** the copy will be **`wonderful.WSencrypted`**. Requires the safer *copy to* button above to be selected.

lemmatising using

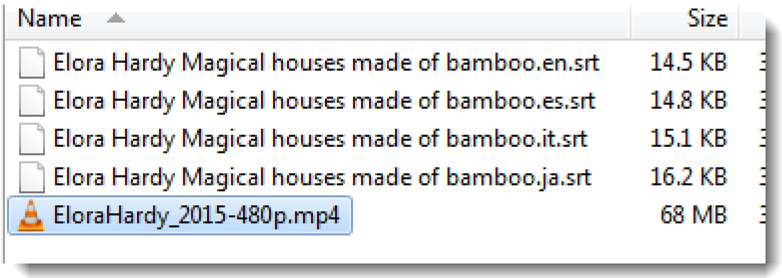
... converts each file using a lemma file 274. If for example your source text has "**`she was tired`**" and your lemma file has **`BE -> AM, WAS, WERE, IS, ARE`**, then you will get "**`she be tired`**" in your converted text file. Where your source text has "**`Was she tired?`**" you'll get "**`Be she tired?`**"

SRT Transcripts

converts SRT files such as those obtained from TED Open Translation Project 215. If using TED files you may need to add some seconds for the standard TED lead-in.

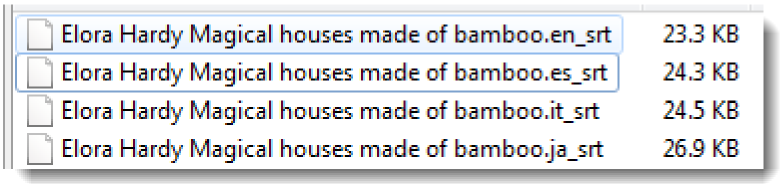
Example

These text files in English (.en), Spanish (.es) Italian (.it) and Japanese (.ja) originally downloaded



Name	Size
Elora Hardy Magical houses made of bamboo.en.srt	14.5 KB
Elora Hardy Magical houses made of bamboo.es.srt	14.8 KB
Elora Hardy Magical houses made of bamboo.it.srt	15.1 KB
Elora Hardy Magical houses made of bamboo.ja.srt	16.2 KB
EloraHardy_2015-480p.mp4	68 MB

got converted thus:



Elora Hardy Magical houses made of bamboo.en_srt	23.3 KB
Elora Hardy Magical houses made of bamboo.es_srt	24.3 KB
Elora Hardy Magical houses made of bamboo.it_srt	24.5 KB
Elora Hardy Magical houses made of bamboo.ja_srt	26.9 KB

To enable Concord to play the .mp4 file 212 I had to change *EloraHardy_2015-480p.mp4* to the same title *Elora Hardy Magical houses made of bamboo.mp4*. Note the file sizes are bigger (converted into Unicode) and the file-names no longer have two dots. This is so that Concord will find a match between its file-name and the transcripts in these 4 languages.

See also: Convert Entire Texts 365

10.7.8 Text Converter: converting BNC XML version

The British National Corpus is a valuable resource but has certain problems as it comes straight off the cdrom:

- it is in Unix format
- it has entities like **`&eacute;`** to represent characters like **é**
- its structure is opaque and file-names mean nothing

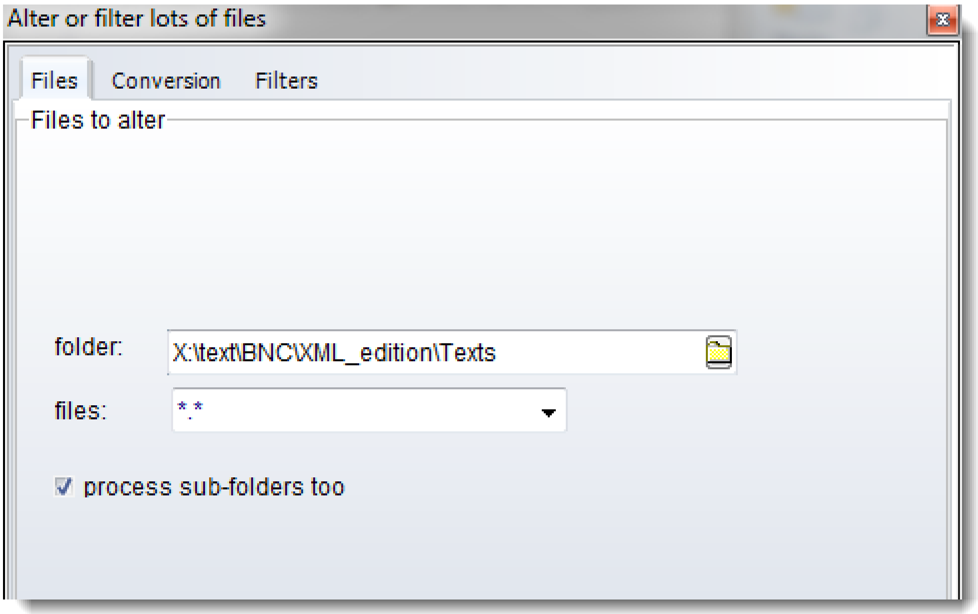
You will find it much easier to use if you

- convert it to Unicode
- filter the files to make a useful structure

as explained at http://lexically.net/wordsmith/Handling_BNC/index.html

The easiest way to do that is in two stages.

Conversion:

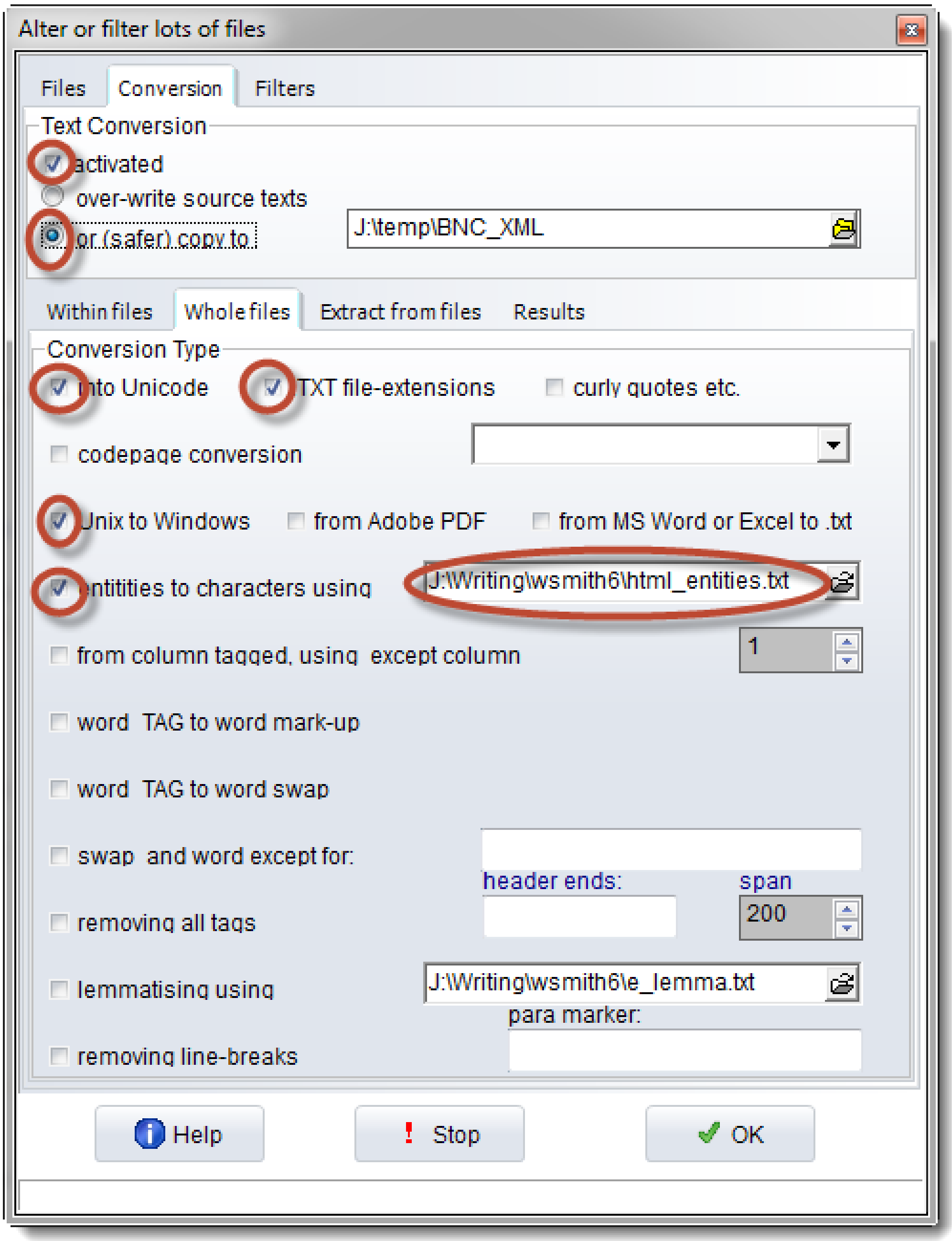


folder: X:\text\BNC\XML_edition\Texts

files: *.*

☑ process sub-folders too

After choosing the texts,



Alter or filter lots of files

Text Conversion

☑ activated

○ over-write source texts

● or (safer) copy to J:\temp\BNC_XML

Conversion Type

☑ into Unicode ☑ TXT file-extensions □ curly quotes etc.

□ codepage conversion

☑ Unix to Windows □ from Adobe PDF □ from MS Word or Excel to .txt

☑ entitities to characters using J:\Writing\wsmith6\html_entities.txt

□ from column tagged, using except column 1

□ word TAG to word mark-up

□ word TAG to word swap

□ swap and word except for:

header ends: span 200

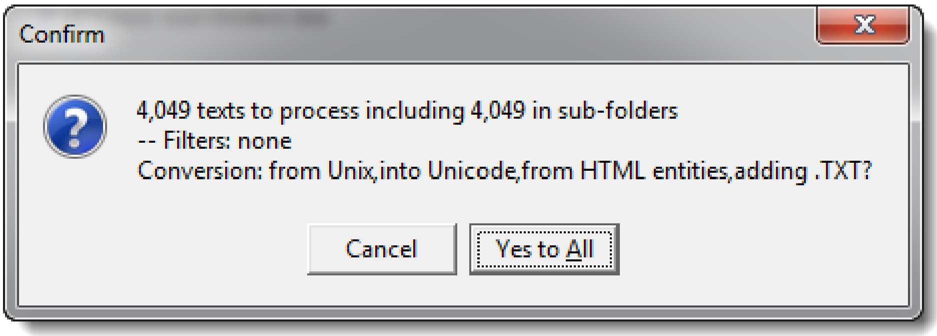
□ removing all tags

□ lemmatising using J:\Writing\wsmith6\e_lemma.txt

para marker:

□ removing line-breaks

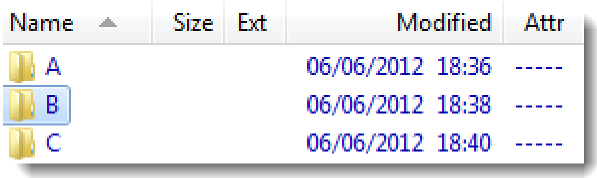
and when you press OK you'll be asked something like this



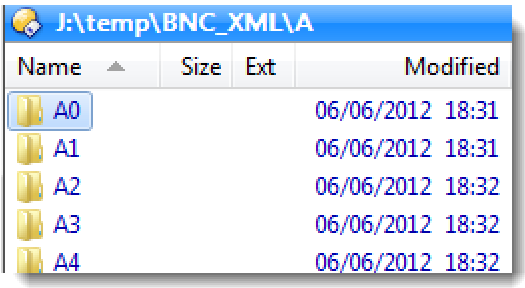
Confirm

4,049 texts to process including 4,049 in sub-folders
-- Filters: none
Conversion: from Unix,into Unicode,from HTML entities,adding .TXT?

After the work is done you will see the BNC texts copied to a similar structure (in our case stemming from `j:\temp`)

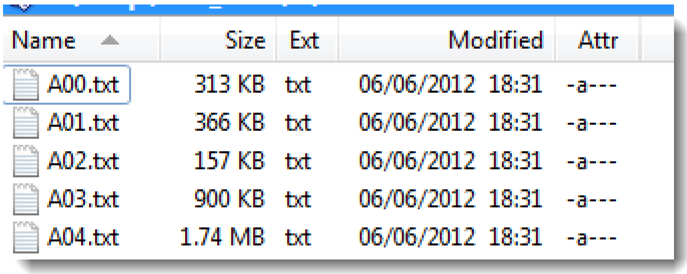


Name	Size	Ext	Modified	Attr
A			06/06/2012 18:36	-----
B			06/06/2012 18:38	-----
C			06/06/2012 18:40	-----



J:\temp\BNC_XML\A

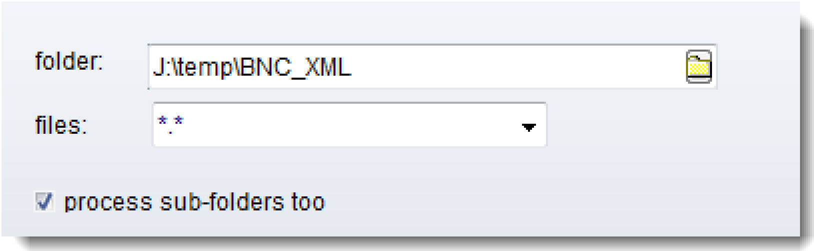
Name	Size	Ext	Modified
A0			06/06/2012 18:31
A1			06/06/2012 18:31
A2			06/06/2012 18:32
A3			06/06/2012 18:32
A4			06/06/2012 18:32



Name	Size	Ext	Modified	Attr
A00.txt	313 KB	txt	06/06/2012 18:31	-a---
A01.txt	366 KB	txt	06/06/2012 18:31	-a---
A02.txt	157 KB	txt	06/06/2012 18:31	-a---
A03.txt	900 KB	txt	06/06/2012 18:31	-a---
A04.txt	1.74 MB	txt	06/06/2012 18:31	-a---

Filter

Choose the converted texts in the first window:

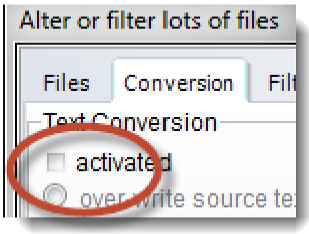


folder: J:\temp\BNC_XML

files: *.*

☑ process sub-folders too

de-activate conversion,



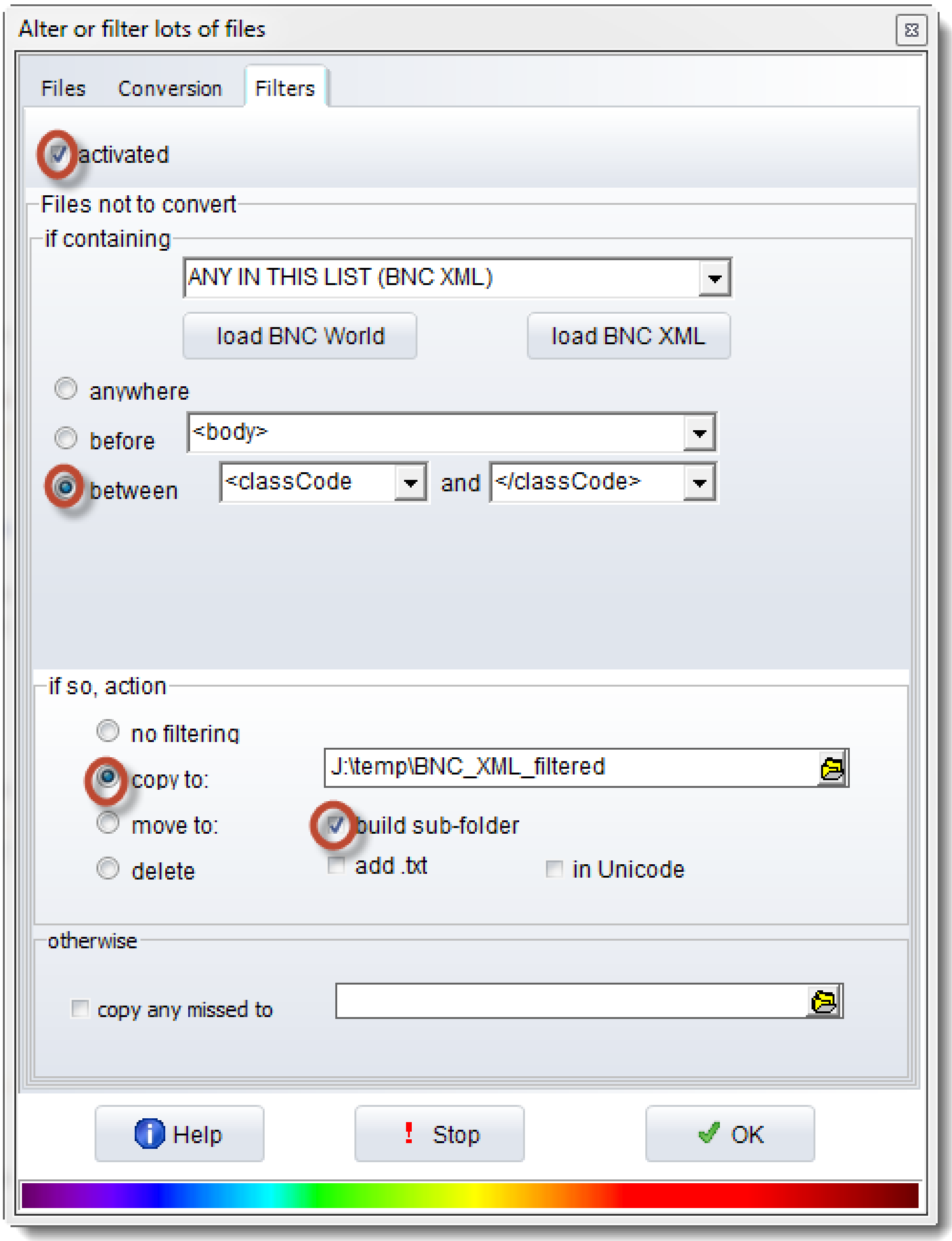
Alter or filter lots of files

Text Conversion

☐ activated

○ over-write source te

and choose filtering like this:



Alter or filter lots of files

☑ activated

Files not to convert

if containing

ANY IN THIS LIST (BNC XML)

○ anywhere

○ before <body>

● between <classCode and </classCode>

if so, action

○ no filtering

● copy to: J:\temp\BNC_XML_filtered

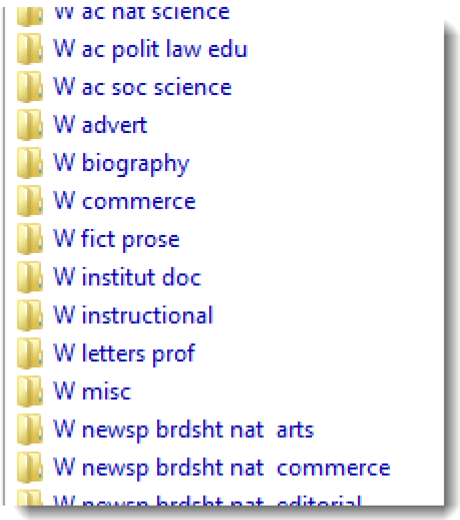
○ move to: ☑ build sub-folder

○ delete □ add .txt □ in Unicode

otherwise

□ copy any missed to

Eventually you should get folder structures like this:



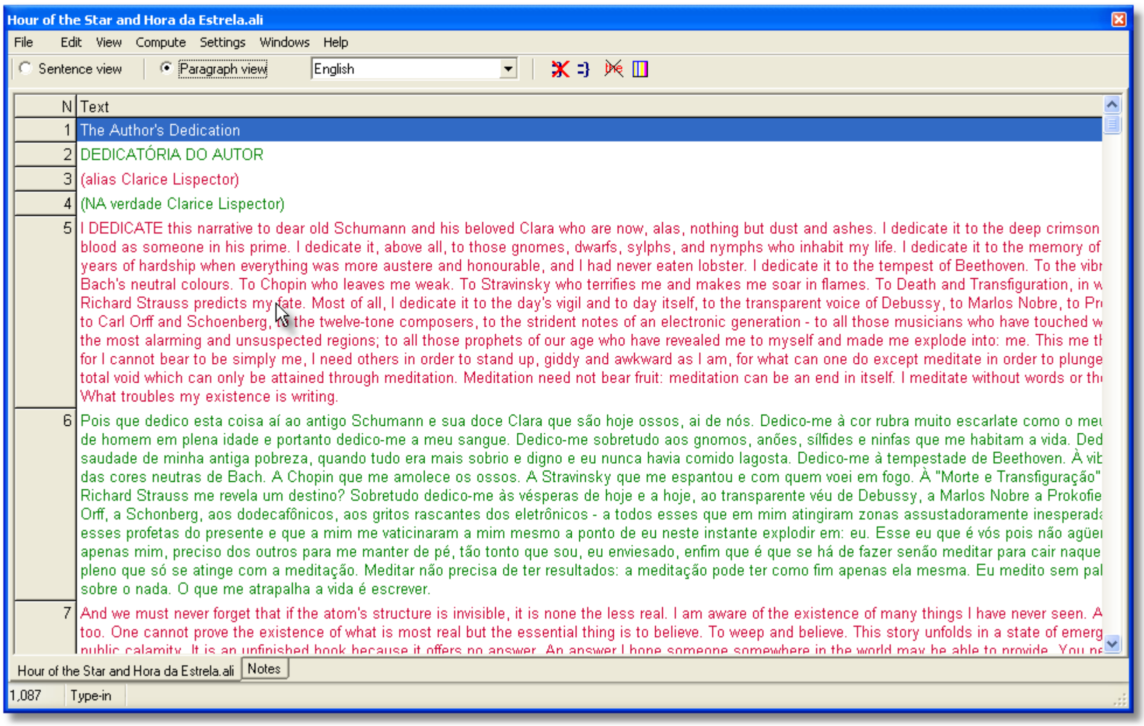
W ac nat science
W ac polit law edu
W ac soc science
W advert
W biography
W commerce
W fict prose
W institut doc
W instructional
W letters prof
W misc
W newsp brdsht nat arts
W newsp brdsht nat commerce
W newsp brdsht nat editorial

10.8 Viewer and Aligner

10.8.1 purpose

This is a program for showing your text or other files, highlighting words of interest. You will see them in plain text format, with tag mark-up shown or hidden as in your tag settings. There are a number of settings 390 and options 387 you can change.

Its main use is to produce an aligned 381 version of 2 or more texts, with alternate sentences or paragraphs from each of them.


Hour of the Star and Hora da Estrela.ali
File Edit View Compute Settings Windows Help
○ Sentence view ● Paragraph view English

N	Text
1	The Author's Dedication
2	DEDICATÓRIA DO AUTOR
3	(alias Clarice Lispector)
4	(NA verdade Clarice Lispector)
5	I DEDICATE this narrative to dear old Schumann and his beloved Clara who are now, alas, nothing but dust and ashes. I dedicate it to the deep crimson blood as someone in his prime. I dedicate it, above all, to those gnomes, dwarfs, sylphs, and nymphs who inhabit my life. I dedicate it to the memory of years of hardship when everything was more austere and honourable, and I had never eaten lobster. I dedicate it to the tempest of Beethoven. To the vibr Bach's neutral colours. To Chopin who leaves me weak. To Stravinsky who terrifies me and makes me soar in flames. To Death and Transfiguration, in w Richard Strauss predicts my fate. Most of all, I dedicate it to the day's vigil and to day itself, to the transparent voice of Debussy, to Marlos Nobre, to Pr to Carl Orff and Schoenberg, to the twelve-tone composers, to the strident notes of an electronic generation - to all those musicians who have touched w the most alarming and unsuspected regions; to all those prophets of our age who have revealed me to myself and made me explode into: me. This me th for I cannot bear to be simply me, I need others in order to stand up, giddy and awkward as I am, for what can one do except meditate in order to plunge total void which can only be attained through meditation. Meditation need not bear fruit: meditation can be an end in itself. I meditate without words or th What troubles my existence is writing.
6	Pois que dedico esta coisa aí ao antigo Schumann e sua doce Clara que são hoje ossos, ai de nós. Dedico-me à cor rubra muito escarlate como o meu de homem em plena idade e portanto dedico-me a meu sangue. Dedico-me sobretudo aos gnomos, anões, sílfides e ninfas que me habitam a vida. Ded saudade de minha antiga pobreza, quando tudo era mais sobrio e digno e eu nunca havia comido lagosta. Dedico-me à tempestade de Beethoven. À vib das cores neutras de Bach. A Chopin que me amolece os ossos. A Stravinsky que me espantou e com quem voei em fogo. À "Morte e Transfiguração" Richard Strauss me revela um destino? Sobretudo dedico-me às vésperas de hoje e a hoje, ao transparente véu de Debussy, a Marlos Nobre a Prokofie Orff, a Schonberg, aos dodecafônicos, aos gritos rascantes dos eletrônicos - a todos esses que em mim atingiram zonas assustadoramente inesperada esses profetas do presente e que a mim me vaticinaram a mim mesmo a ponto de eu neste instante explodir em: eu. Esse eu que é vós pois não agüe apenas mim, preciso dos outros para me manter de pé, tão tonto que sou, eu enviesado, enfim que é que se há de fazer senão meditar para cair naque pleno que só se atinge com a meditação. Meditar não precisa de ter resultados: a meditação pode ter como fim apenas ela mesma. Eu medito sem pal sobre o nada. O que me atrapalha a vida é escrever.
7	And we must never forget that if the atom's structure is invisible, it is none the less real. I am aware of the existence of many things I have never seen. A too. One cannot prove the existence of what is most real but the essential thing is to believe. To weep and believe. This story unfolds in a state of emerg public calamity. It is an unfinished book because it offers no answer. An answer I hope someone somewhere in the world may be able to provide. You ne

See also: Viewer & Aligner settings 390, Viewer & Aligner options 387, an example of aligning 381

10.8.2 index

Explanations

What is the Viewer & Aligner and what's it for? 379
an example of aligning 381
Settings 390
Viewing Options 387
What to do if it doesn't do what I want... 455
Searching for Short Sentences 392
Joining/Splitting 389
Aligning a Dual Text 384
Finding translation mis-matches 391
The technical side... 390

see also : WordSmith Main Index 2

10.8.3 aligning with Viewer & Aligner

This feature aligns the sentences in two files. Translators need to study differences between an original and a translation. Other linguists might want it to study differences between two versions of a text in the same language. Students of different languages 81 can use it as they might use dual language readings, to study closely the differences e.g. in word order.
It helps you produce a new text which consists of the two files, with sentences interspersed. That way you can compare the translation with the original.

Example

Original : *Der Knabe sagte diesen Gedanken dem Schwesterchen, und diese folgte. Allein auch der Weg auf den Hals hinab war nicht zu finden. So klar die Sonne schien, …(from Stifter's Bergkristall, translated by Harry Steinhauer, in German Stories, Bantam Books 1961)*
Translation: *The boy communicated this thought to his sister and she followed him. But the road down the neck could not be found either. Though the sun shone clearly, …*

Aligned text:
<G1> Der Knabe sagte diesen Gedanken dem Schwesterchen, und diese folgte.
<E1> The boy communicated this thought to his sister and she followed him.
<G2> Allein auch der Weg auf den Hals hinab war nicht zu finden.
<E2> But the road down the neck could not be found either.
<G3> So klar die Sonne schien, …
<E3> Though the sun shone clearly, …

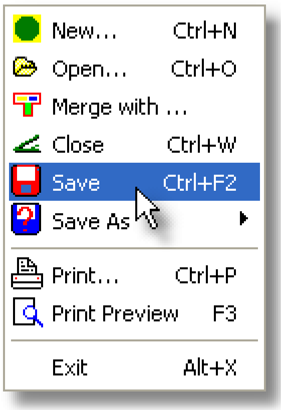
An aligned text like this helps you identify additions and omissions, normalisations, style changes, word order preferences. In this case the translator has chosen to avoid very close equivalence.

See also: an example of aligning 381, Aligning and moving 384

10.8.4 example of aligning

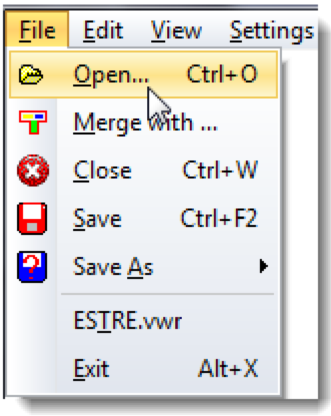
How to do it -- a Portuguese and English example

1. Read in 387 your Portuguese text (eg. **`Hora da Estrela.TXT`**), and checking its sentences and paragraphs break 389 the way you like. Try "Unusual Lines 392" to help identify oddities.
2. Save it

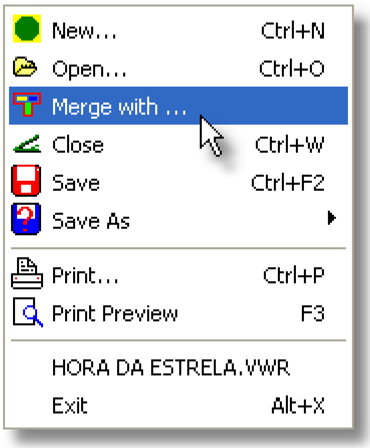


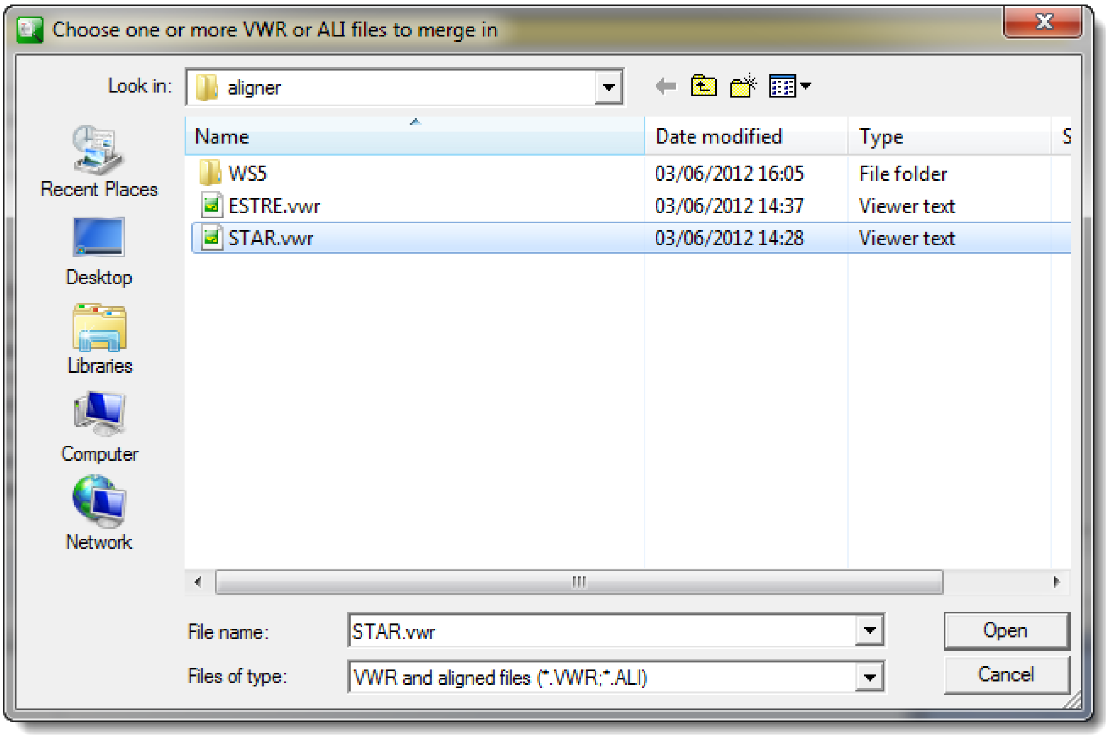
and it will (by default) get your filename**.VWR,** eg. **Hora da Estrela.VWR.**
(It is important to do that, as a .VWR file knows the language, colour settings etc. and the cleaning up work you've done, whereas the .TXT file is just the original text file you read in.)

3. Do the same steps 1 and 2 for your English text -- you will now have e.g. **Hour of the Star.VWR**.
4. You could if desired repeat with the Spanish -- **Hora de la Estrella.txt** giving **Hora de la Estrella.VWR**, (or German, Russian, Arabic, etc.).
5. Now open your Portuguese **Hora da Estrela.VWR**



6. and then *File* | *Merge*



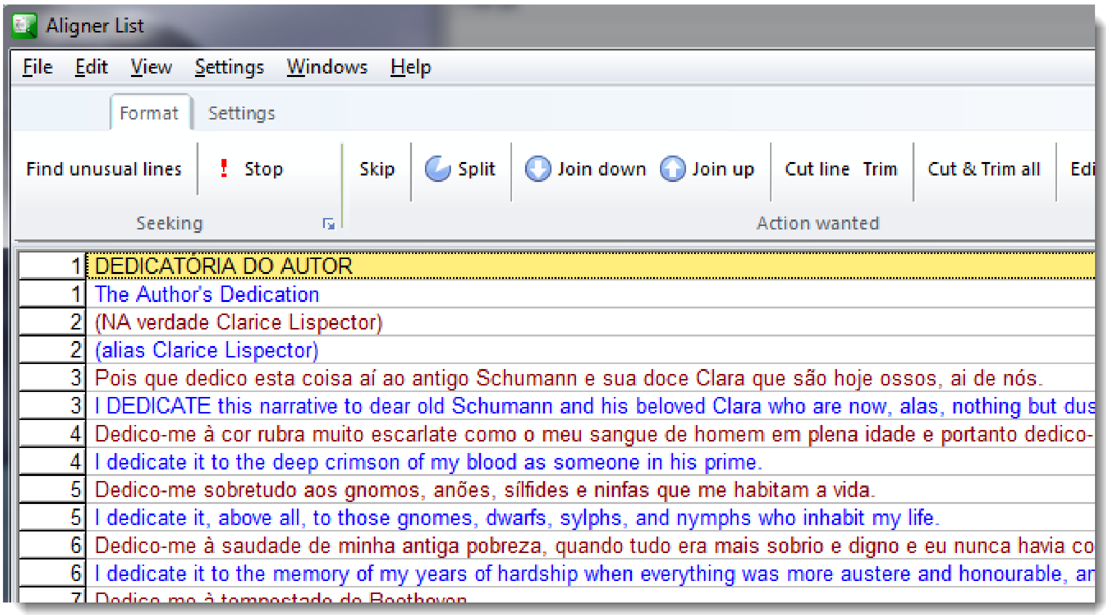

Choose one or more VWR or ALI files to merge in

Look in: aligner

Name	Date modified	Type	S
WS5	03/06/2012 16:05	File folder	
ESTRE.vwr	03/06/2012 14:37	Viewer text	
STAR.vwr	03/06/2012 14:28	Viewer text	

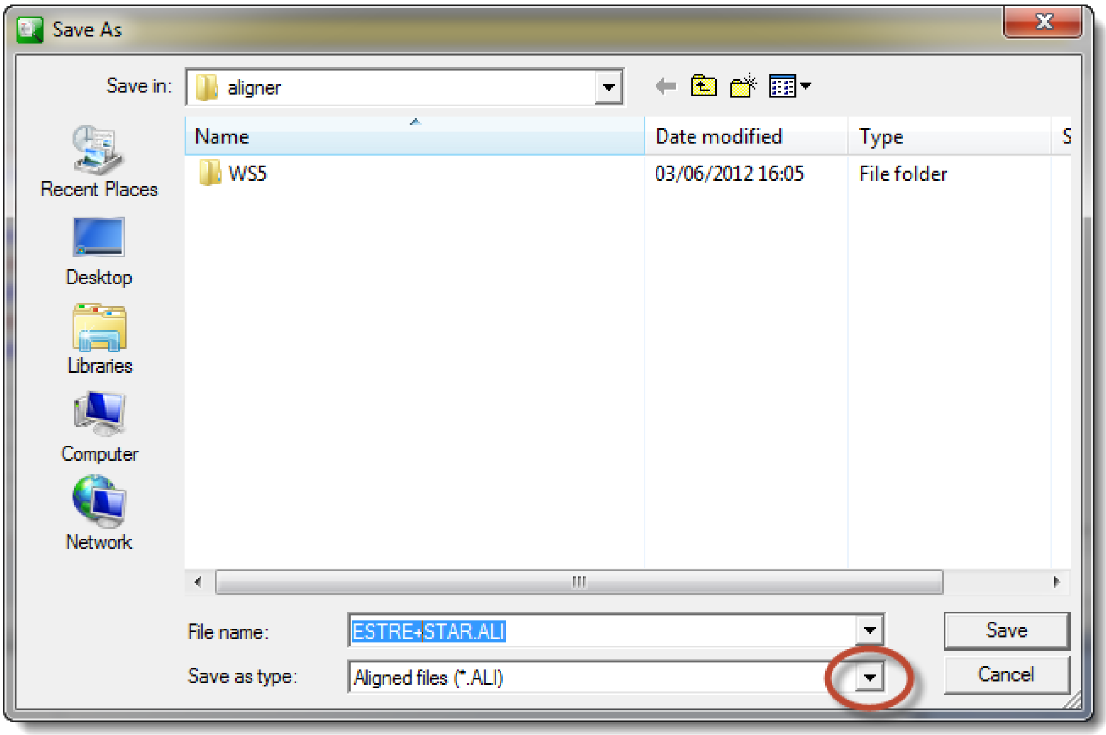
File name: STAR.vwr

Files of type: VWR and aligned files (*.VWR;*.ALI)



1 DEDICATÓRIA DO AUTOR
1 The Author's Dedication
2 (NA verdade Clarice Lispector)
2 (alias Clarice Lispector)
3 Pois que dedico esta coisa aí ao antigo Schumann e sua doce Clara que são hoje ossos, ai de nós.
3 I DEDICATE this narrative to dear old Schumann and his beloved Clara who are now, alas, nothing but dus
4 Dedico-me à cor rubra muito escarlate como o meu sangue de homem em plena idade e portanto dedico-
4 I dedicate it to the deep crimson of my blood as someone in his prime.
5 Dedico-me sobretudo aos gnomos, anões, sílfides e ninfas que me habitam a vida.
5 I dedicate it, above all, to those gnomes, dwarfs, sylphs, and nymphs who inhabit my life.
6 Dedico-me à saudade de minha antiga pobreza, quando tudo era mais sobrio e digno e eu nunca havia co
6 I dedicate it to the memory of my years of hardship when everything was more austere and honourable, ar
7 Dedico-me à tempestade de Beethoven

7. Finally *File* | *Save choosing Aligned files* (**`.ALI`**) as the format.



Save in: aligner

WS5 03/06/2012 16:05 File folder

File name: ESTRE+STAR.ALI

Save as type: Aligned files (*.ALI)

10.8.5 aligning and moving

You may well want to alter sentence ordering. The translator may have used three sentences where the original had only one. You can also merge paragraphs.

adjusting by dragging with the mouse

To merge sentences or paragraphs, simply grab and drag it up to the next one above in the same language. Or use the Join button. Or press F4.
To split a sentence or paragraph, choose the Split button or press Ctrl+F4.

Finally you will want to save (Ctrl+F2) the results 101.

See also: Viewer & Aligner contents 380

10.8.6 editing

While Viewer & Aligner is not a full word-processor, some editing facilities have been built in to help deal with common formatting problems:

- Split: allows you to choose where a line should be divided in two.
- Join down , Join up: these buttons merge a line with another one. You can achieve this also by simply dragging.
- Cut line: removes any blank lines.
- Trim: this goes through each sentence of the text, removing any redundant spaces -- where there are two or more consecutive spaces they will be reduced to one.
- Cut & Trim All does these actions on the whole text.
- Edit opens up a window allowing you to edit the whole of the current sentence or paragraph.
- Heading: allows you to treat a line as a heading, and if so makes it look bold.
- Find unusual lines 392: this identifies cases where a sentence or paragraph does not start with a capital letter or number -- you will probably want to join 389 it to the one above, or where a line is unusually short, etc.
- Find short lines 392

You will then want to save (Ctrl+F2) your text.

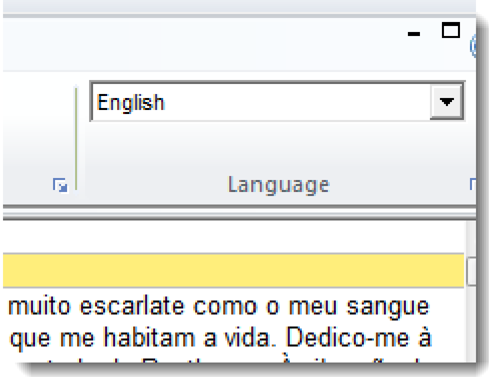
You can also:

- open a new file for viewing (you can open any number of text files within Viewer & Aligner)
- copy a text file to the clipboard 422 (select, then press Control+Ins)
- print the whole or part of the currently active text file
- search for words or phrases (press F12)

10.8.7 languages

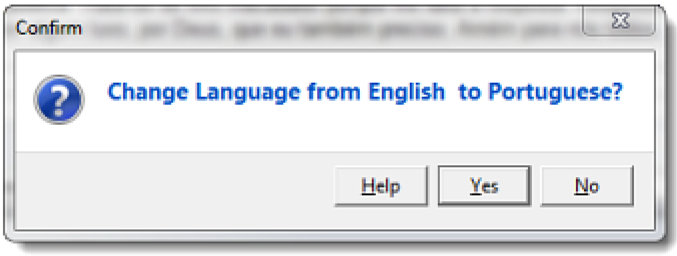
Each Viewer file (**.VWR**) has its own language. Each Aligner file (**.ALI**) has one language for each of the component sections. (They could all be the same, if for example you were analysing various different editions of a Shakespeare play they'd all be English.) The set of languages available is that defined using the Languages Chooser 83.

You can change the language to one of your previously defined languages 84 using the drop-down list. Here is an example where a Portuguese language plain TXT text file was opened and the default language was English.



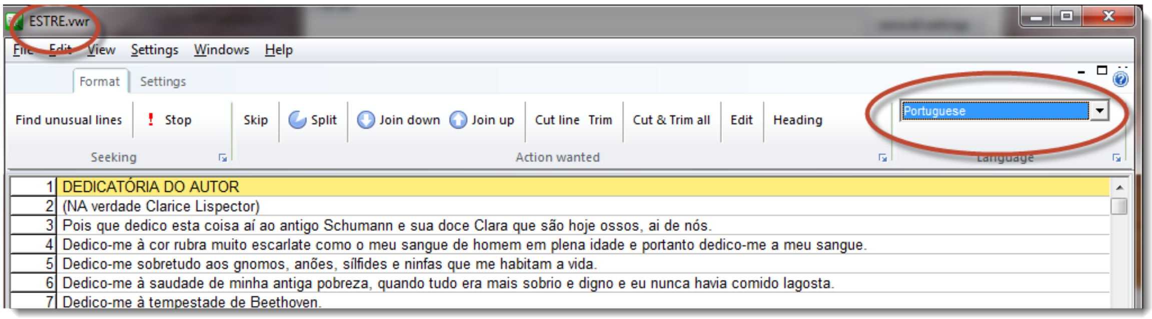
muito escarlate como o meu sangue
que me habitam a vida. Dedico-me à

When Portuguese was chosen in the drop-down list, and



Change Language from English to Portuguese?

agreed to, it was possible to save the result (as a .VWR file) so that henceforth it would know which language to use.



1	DEDICATÓRIA DO AUTOR
2	(NA verdade Clarice Lispector)
3	Pois que dedico esta coisa aí ao antigo Schumann e sua doce Clara que são hoje ossos, ai de nós.
4	Dedico-me à cor rubra muito escarlate como o meu sangue de homem em plena idade e portanto dedico-me a meu sangue.
5	Dedico-me sobretudo aos gnomos, anões, sílfides e ninfas que me habitam a vida.
6	Dedico-me à saudade de minha antiga pobreza, quando tudo era mais sobrio e digno e eu nunca havia comido lagosta.
7	Dedico-me à tempestade de Beethoven.

10.8.8 numbering sentences & paragraphs

You can use the **Viewer & Aligner** to make a copy of your text with all the sentences and/or paragraphs tagged with **<S> and <P>**.
To do this, simply read in the text file in, choose *Edit* | *Insert Tags*, then save it as a text file 102.

See also: Viewer & Aligner contents 380

10.8.9 options

Mode: Sentence/Paragraph

This switches between Sentence mode and Paragraph mode. In other words you can choose to view your text files with each row of the display taking up a sentence or a paragraph.

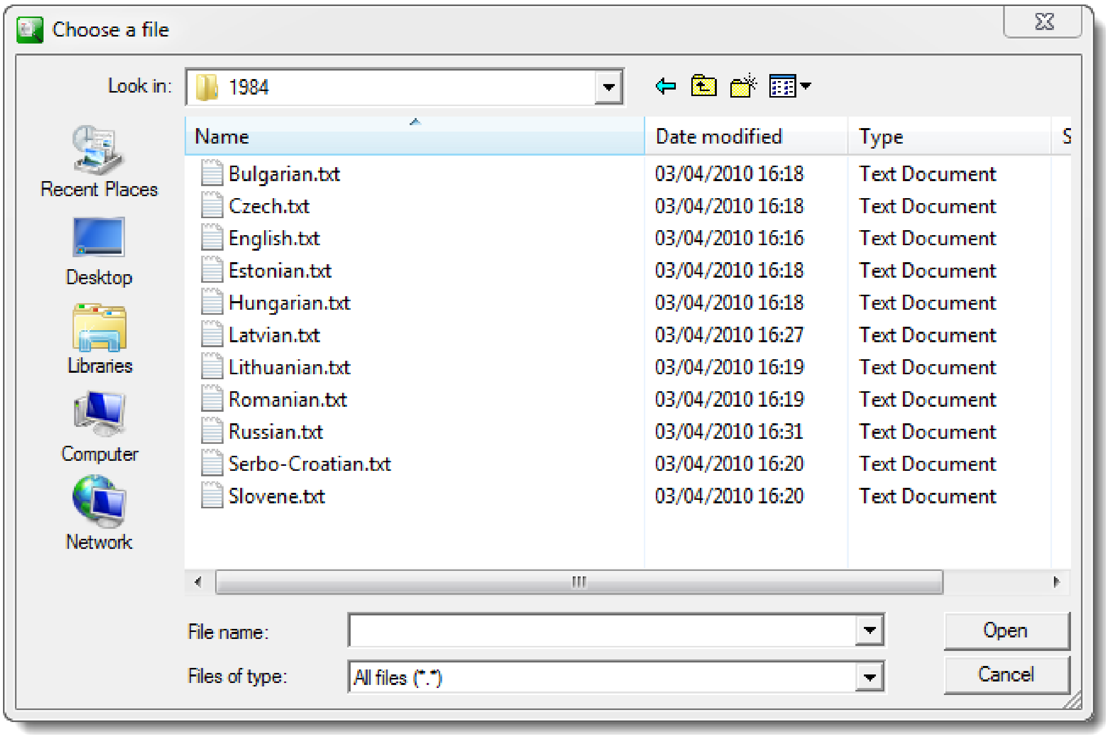
Likewise, you can make an dual aligned text by interspersing either paragraphs or sentences. The other functions (e.g. joining, splitting 389) work in the same way in either mode.

Colours

The various texts in your aligned text will have different colours associated with them. Colours can be changed using the ▥ button.

10.8.10 reading in a plain text

In Viewer and Aligner, choose *File* | *Open,* and select your plain text file.



Choose a file

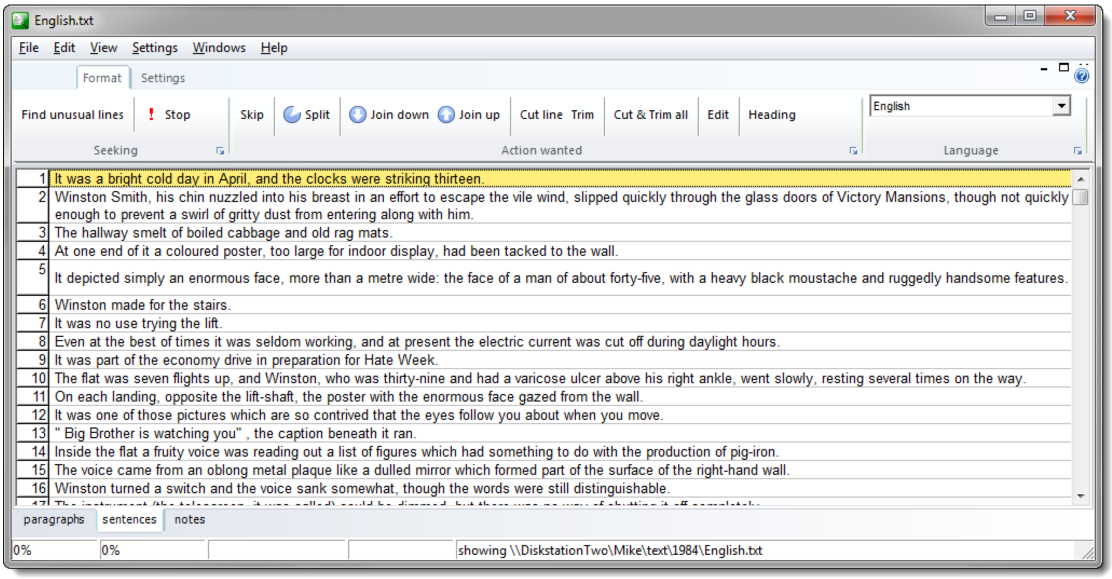
Look in: 1984

Name	Date modified	Type
Bulgarian.txt	03/04/2010 16:18	Text Document
Czech.txt	03/04/2010 16:18	Text Document
English.txt	03/04/2010 16:16	Text Document
Estonian.txt	03/04/2010 16:18	Text Document
Hungarian.txt	03/04/2010 16:18	Text Document
Latvian.txt	03/04/2010 16:27	Text Document
Lithuanian.txt	03/04/2010 16:19	Text Document
Romanian.txt	03/04/2010 16:19	Text Document
Russian.txt	03/04/2010 16:31	Text Document
Serbo-Croatian.txt	03/04/2010 16:20	Text Document
Slovene.txt	03/04/2010 16:20	Text Document

File name:

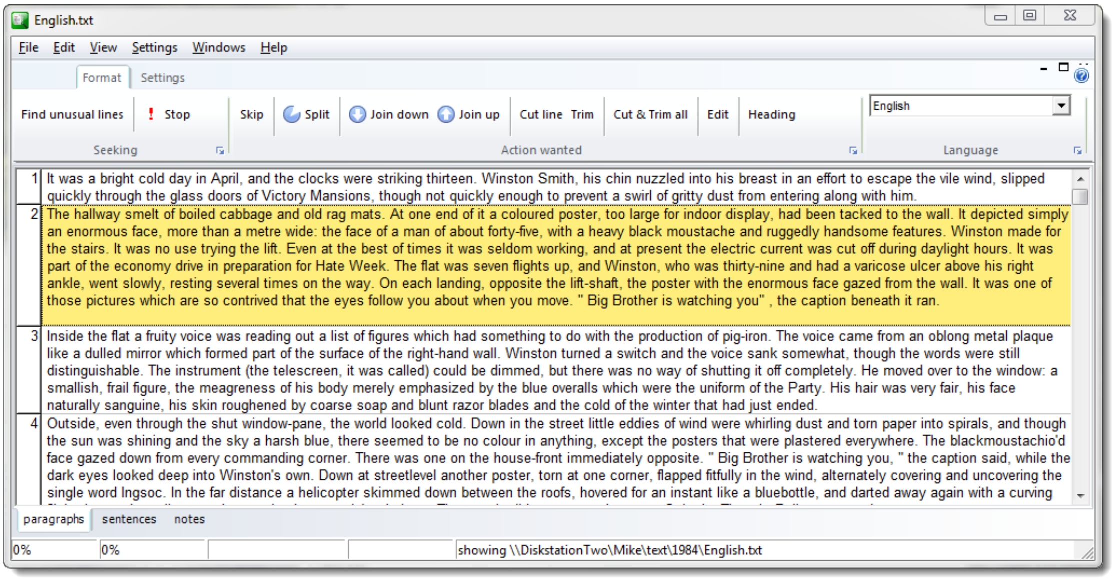
Files of type: All files (*.*)

and you may see this sort of thing in *Sentence view*,



1	It was a bright cold day in April, and the clocks were striking thirteen.
2	Winston Smith, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of Victory Mansions, though not quickly enough to prevent a swirl of gritty dust from entering along with him.
3	The hallway smelt of boiled cabbage and old rag mats.
4	At one end of it a coloured poster, too large for indoor display, had been tacked to the wall.
5	It depicted simply an enormous face, more than a metre wide: the face of a man of about forty-five, with a heavy black moustache and ruggedly handsome features.
6	Winston made for the stairs.
7	It was no use trying the lift.
8	Even at the best of times it was seldom working, and at present the electric current was cut off during daylight hours.
9	It was part of the economy drive in preparation for Hate Week.
10	The flat was seven flights up, and Winston, who was thirty-nine and had a varicose ulcer above his right ankle, went slowly, resting several times on the way.
11	On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall.
12	It was one of those pictures which are so contrived that the eyes follow you about when you move.
13	" Big Brother is watching you" , the caption beneath it ran.
14	Inside the flat a fruity voice was reading out a list of figures which had something to do with the production of pig-iron.
15	The voice came from an oblong metal plaque like a dulled mirror which formed part of the surface of the right-hand wall.
16	Winston turned a switch and the voice sank somewhat, though the words were still distinguishable.

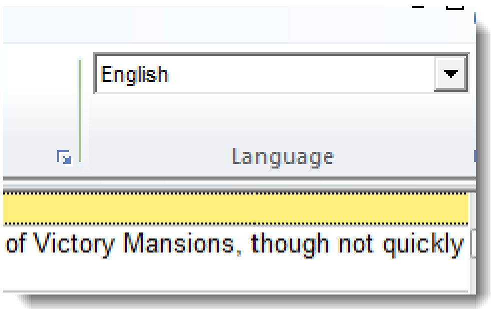
or in *Paragraph view*,



1	It was a bright cold day in April, and the clocks were striking thirteen. Winston Smith, his chin nuzzled into his breast in an effort to escape the vile wind, slipped quickly through the glass doors of Victory Mansions, though not quickly enough to prevent a swirl of gritty dust from entering along with him.
2	The hallway smelt of boiled cabbage and old rag mats. At one end of it a coloured poster, too large for indoor display, had been tacked to the wall. It depicted simply an enormous face, more than a metre wide: the face of a man of about forty-five, with a heavy black moustache and ruggedly handsome features. Winston made for the stairs. It was no use trying the lift. Even at the best of times it was seldom working, and at present the electric current was cut off during daylight hours. It was part of the economy drive in preparation for Hate Week. The flat was seven flights up, and Winston, who was thirty-nine and had a varicose ulcer above his right ankle, went slowly, resting several times on the way. On each landing, opposite the lift-shaft, the poster with the enormous face gazed from the wall. It was one of those pictures which are so contrived that the eyes follow you about when you move. " Big Brother is watching you" , the caption beneath it ran.
3	Inside the flat a fruity voice was reading out a list of figures which had something to do with the production of pig-iron. The voice came from an oblong metal plaque like a dulled mirror which formed part of the surface of the right-hand wall. Winston turned a switch and the voice sank somewhat, though the words were still distinguishable. The instrument (the telescreen, it was called) could be dimmed, but there was no way of shutting it off completely. He moved over to the window: a smallish, frail figure, the meagreness of his body merely emphasized by the blue overalls which were the uniform of the Party. His hair was very fair, his face naturally sanguine, his skin roughened by coarse soap and blunt razor blades and the cold of the winter that had just ended.
4	Outside, even through the shut window-pane, the world looked cold. Down in the street little eddies of wind were whirling dust and torn paper into spirals, and though the sun was shining and the sky a harsh blue, there seemed to be no colour in anything, except the posters that were plastered everywhere. The blackmoustachio'd face gazed down from every commanding corner. There was one on the house-front immediately opposite. " Big Brother is watching you, " the caption said, while the dark eyes looked deep into Winston's own. Down at streetlevel another poster, torn at one corner, flapped fitfully in the wind, alternately covering and uncovering the single word Ingsoc. In the far distance a helicopter skimmed down between the roofs, hovered for an instant like a bluebottle, and darted away again with a curving

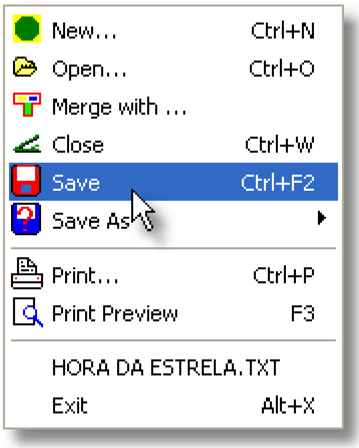
Edit it, as necessary, e.g. splitting or merging 389 paragraphs or sentences. There's a taskbar with buttons to help above the text.

Ensure the language is right:



of Victory Mansions, though not quickly

And save it as a **.VWR** file:



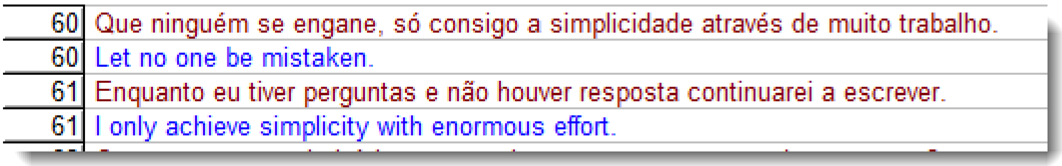
See also: example of aligning 381

10.8.11 joining and splitting

Joining

The easiest way to join two sentences is simply to drag the one you want to move onto its neighbour above. Or select the lower of the two and press F4 or use the button ( )

In this example, sentence 60 in Portuguese got represented as two sentences, 60 and 61, in English.

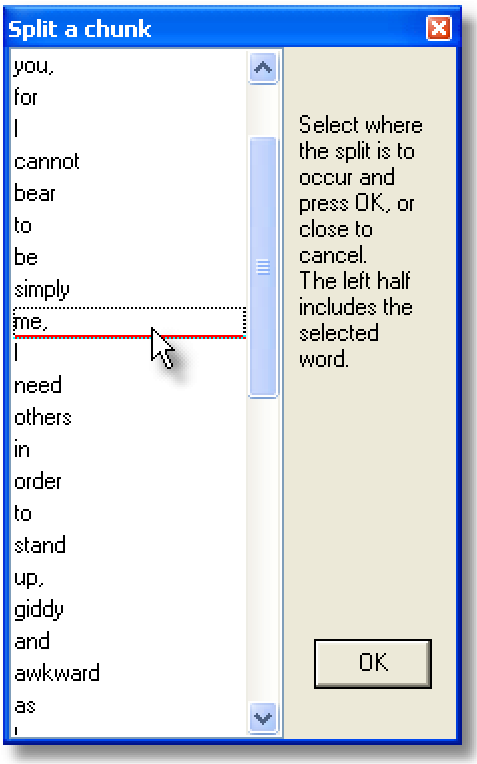


60	Que ninguém se engane, só consigo a simplicidade através de muito trabalho.
60	Let no one be mistaken.
61	Enquanto eu tiver perguntas e não houver resposta continuarei a escrever.
61	I only achieve simplicity with enormous effort.

Splitting in two

To split a sentence, press . You will get a list of the words. Click on the word which should end the sentence, then press OK.

example



you,
for
I
cannot
bear
to
be
simply
me,
I
need
others
in
order
to
stand
up,
giddy
and
awkward
as

Select where the split is to occur and press OK, or close to cancel. The left half includes the selected word.

This will insert the words which follow (**`I need others`** etc.) into a new line below.

See also: Viewer & Aligner contents 380

10.8.12 settings

1. What constitutes a "short" sentence or paragraph (default: less than 25 characters)
2. Whether you want to do a lower-case check when Finding Unusual Lines

The settings are standard ones found in most of the Tools:

Colours 60
Font 78
Printing 80
Text Characteristics 124
Review all Settings 113

10.8.13 technical aspects

When is a sentence not a sentence?

There is no perfect mechanical way of determining sentence-breaks. For example, a heading may well have no final full stop but would normally not be considered part of the sentence which follows it. And a sentence may often have no final full stop, if what follows it is a list of items.

The algorithm used by Viewer & Aligner is: a sentence ends when it meets the requirements explained in the definition of a sentence 426. The same routine is used as in WordList.

Consider this chunk from *A Tale of Two Cities*:

"Wo-ho!" said the coachman. "So, then! One more pull and you're at the top and be damned to you, for I have had trouble enough to get you to it! - Joe!"

Viewer & Aligner will mistakenly consider - **Joe!** as a separate sentence, but handles **"Wo-ho!" said the coachman.** as one: though the program would split it in two if the word after **ho!** had a capital lettter (e.g. in **Wild Bill, the coachman, said.**)

Viewer & Aligner cannot therefore be expected to handle all sentence boundaries exactly as *you* would. (**I saw Mr. Smith.** would be considered two sentences; several headings may be bundled together as one sentence.) For this reason you can choose *Find Short Sentences* to seek out 392 any odd one-word sentences.

See also: Viewer & Aligner contents 380

10.8.14 translation mis-matches

Viewer & Aligner can help find cases where alignment has slipped (one sentence having been translated as two or three). One method is to use the menu item *Match by Capitals*. This searches for matching proper nouns in the two versions: if say **Paris** is mentioned in sentences 25 of the source text and not in sentence 25 of the translation but in sentence 27, it is very likely that some slippage has occurred.

Viewer & Aligner will search forwards from the current text sentence on, and will tell you where there's a mis-match. You should then search back from that point to find where the sentences start to diverge. It may be useful to sample every 10 or every 20 to speed up the search for slippage. When you find the problem, un-join 389 or join 389 and/or edit the text as appropriate, then save it.

See also: The technical side... 390, Finding unusual sentences 392, Viewer & Aligner contents 380

10.8.15 troubleshooting

Can't see the whole sentence or paragraph

Press ⇕ to "auto-size" the lines in your display. This adjusts line heights according to the current highlighted column of data.

Can't see the whole text file

Press ▭ to "refresh" the display.

Don't like the colours

Change colours using . The colours initially used for each language version in the dual-language window are the same colours as used for primary sorting and secondary sorting in **Concord**.

See also: Viewer & Aligner contents 380

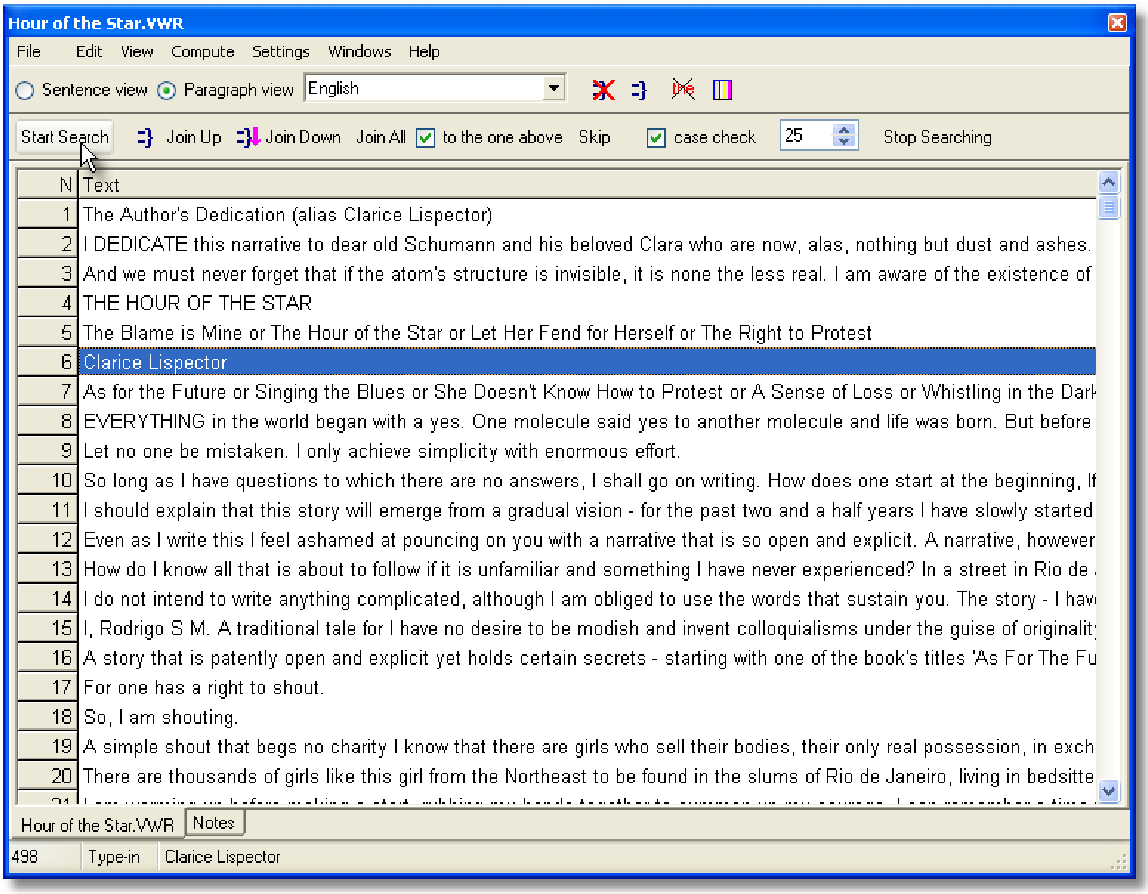
10.8.16 unusual lines

It can be useful to seek unusually short sentences to see whether your originals have been handled as you want. Because Viewer & Aligner uses full stops, question marks and exclamation marks as sentence-boundary indicators, you will find a string like "Hello! Paul! Come here!" is broken into 3 very short sentences. Depending on your purposes you may wish to consider these as one sentence, e.g. if a translator has translated them as one ("Oi, Paulo, venha cá!") .

This function can also find lower-case lines: where a sentence or paragraph does not start with a capital letter or number -- you will probably want to join it to the one above. This problem is common if the text has been saved as "text only with line breaks" (where an <Enter> comes at the end of each line whether or not it is the end of a paragraph.)

Seeking

Use the Find Unusual Toolbar menu item (??) and then press *Start Search*. Viewer & Aligner will go to the next possibly problematic sentence or paragraph and you will probably want to join 389 it by pressing Join Up (to the one above), Join Down, or Skip.



○ Sentence view ● Paragraph view English

☑ to the one above ☑ case check 25

N	Text
1	The Author's Dedication (alias Clarice Lispector)
2	I DEDICATE this narrative to dear old Schumann and his beloved Clara who are now, alas, nothing but dust and ashes.
3	And we must never forget that if the atom's structure is invisible, it is none the less real. I am aware of the existence of
4	THE HOUR OF THE STAR
5	The Blame is Mine or The Hour of the Star or Let Her Fend for Herself or The Right to Protest
6	Clarice Lispector
7	As for the Future or Singing the Blues or She Doesn't Know How to Protest or A Sense of Loss or Whistling in the Dark
8	EVERYTHING in the world began with a yes. One molecule said yes to another molecule and life was born. But before
9	Let no one be mistaken. I only achieve simplicity with enormous effort.
10	So long as I have questions to which there are no answers, I shall go on writing. How does one start at the beginning, if
11	I should explain that this story will emerge from a gradual vision - for the past two and a half years I have slowly started
12	Even as I write this I feel ashamed at pouncing on you with a narrative that is so open and explicit. A narrative, however
13	How do I know all that is about to follow if it is unfamiliar and something I have never experienced? In a street in Rio de .
14	I do not intend to write anything complicated, although I am obliged to use the words that sustain you. The story - I have
15	I, Rodrigo S M. A traditional tale for I have no desire to be modish and invent colloquialisms under the guise of originality
16	A story that is patently open and explicit yet holds certain secrets - starting with one of the book's titles 'As For The Fu
17	For one has a right to shout.
18	So, I am shouting.
19	A simple shout that begs no charity I know that there are girls who sell their bodies, their only real possession, in exch
20	There are thousands of girls like this girl from the Northeast to be found in the slums of Rio de Janeiro, living in bedsitte

498 Type-in Clarice Lispector

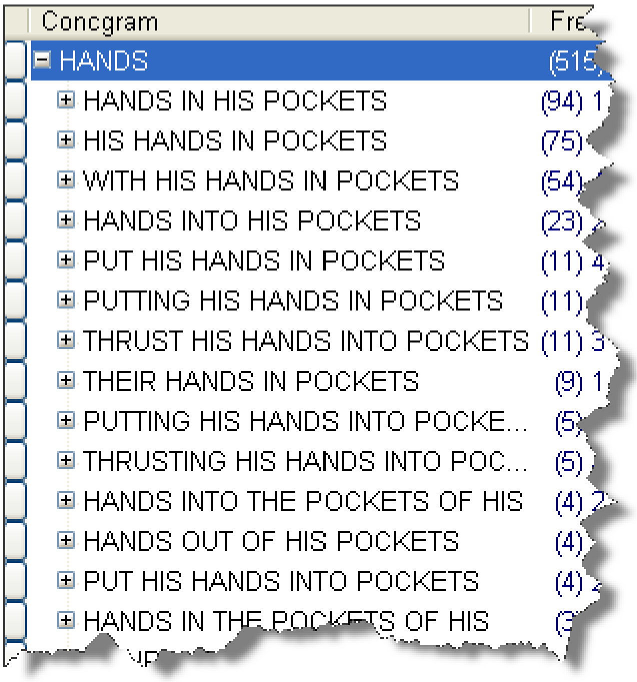
"Case check" switches on or off the search for lower-case sentence starts. The number (25 in the example above) is for you to determine the number of characters counting as a short sentence or paragraph.

See also: Settings 390, The technical side... 390, Finding translation mis-matches 391, Viewer & Aligner contents 380

10.9 WSConcGram

10.9.1 aims

A program for finding concgrams 394, essentially related pairs, triplets, quadruplets (etc.) of words which are related.



Concgram	Fre
⊟ HANDS	(515
⊞ HANDS IN HIS POCKETS	(94) 1
⊞ HIS HANDS IN POCKETS	(75)
⊞ WITH HIS HANDS IN POCKETS	(54)
⊞ HANDS INTO HIS POCKETS	(23)
⊞ PUT HIS HANDS IN POCKETS	(11) 4
⊞ PUTTING HIS HANDS IN POCKETS	(11)
⊞ THRUST HIS HANDS INTO POCKETS	(11) 3
⊞ THEIR HANDS IN POCKETS	(9) 1
⊞ PUTTING HIS HANDS INTO POCKE...	(5)
⊞ THRUSTING HIS HANDS INTO POC...	(5)
⊞ HANDS INTO THE POCKETS OF HIS	(4) 2
⊞ HANDS OUT OF HIS POCKETS	(4)
⊞ PUT HIS HANDS INTO POCKETS	(4)
⊞ HANDS IN THE POCKETS OF HIS	(3

See also : definition of concgram 394, settings 395, running WSConcGram 395, filtering 402, viewing the output 397.

10.9.2 definition of a concgram

For years it has been easy to search for or identify consecutive clusters (n-grams) such as **AT THE END OF, MERRY CHRISTMAS** or **TERM TIME**. It has also been possible to find non-consecutive linkages such as **STRONG** within the horizons 180 of **TEA** by adapting searches to find context words 202. The concgram procedure takes a whole corpus of text and finds all sorts of combinations like the ones above, whether consecutive or not.

Cheng, Greaves & Warren (2006:414) define a concgram like this

> For our purposes, a ‘concgram’ is all of the permutations of constituency variation and positional variation generated by the association of two or more words. This means that the associated words comprising a particular concgram may be the source of a number of ‘collocational patterns’ (Sinclair 2004:xxvii). In fact, the hunt for what we term ‘concgrams’ has a fairly long history dating back to the 1980s (Sinclair 2005, personal communication) when the Cobuild team at the University of Birmingham led by Professor John Sinclair attempted, with limited success, to devise the means to automatically search for non-contiguous sequences of associated words.

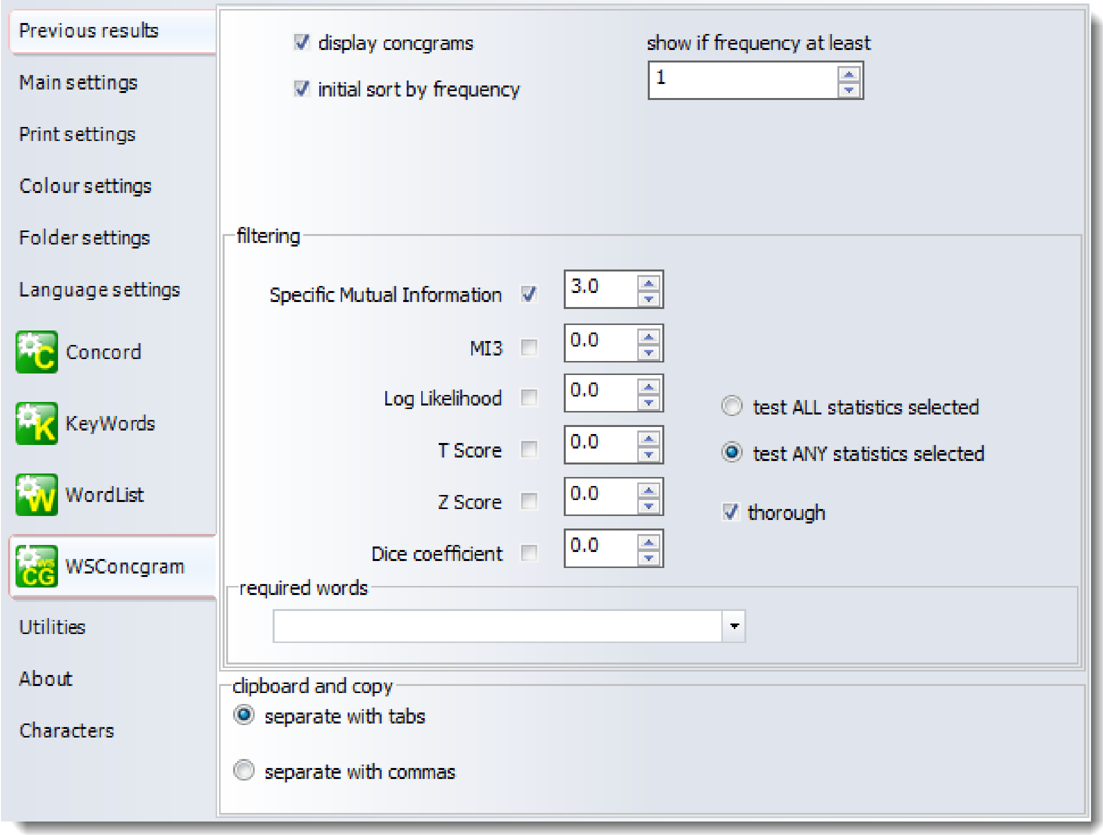
Essentially what they were seeking in developing the ConcGram (©) program was "a search-engine, which on top of the capability to handle constituency variation (i.e. AB, ACB), also handles positional variation (i.e. AB, BA), conducts fully automated searches, and searches for word associations of any size." (2006:413)

WSConcGram is developed in homage to this idea.

See also: bibliography 417, settings 395, running WSConcGram 395, filtering 402, viewing the output 397.

10.9.3 settings

The settings are found in the main Controller.



☑ display concgrams

☑ initial sort by frequency

show if frequency at least: 1

filtering

Specific Mutual Information	☑	3.0
MI3	□	0.0
Log Likelihood	□	0.0
T Score	□	0.0
Z Score	□	0.0
Dice coefficient	□	0.0

○ test ALL statistics selected

● test ANY statistics selected

☑ thorough

required words

clipboard and copy

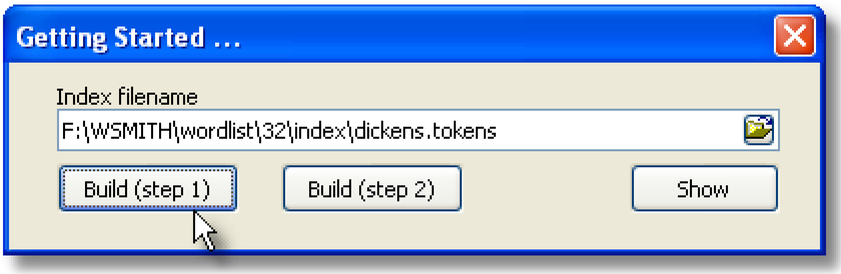
● separate with tabs

○ separate with commas

10.9.4 generating concgrams

To start, as usual, choose *File* | *New*.

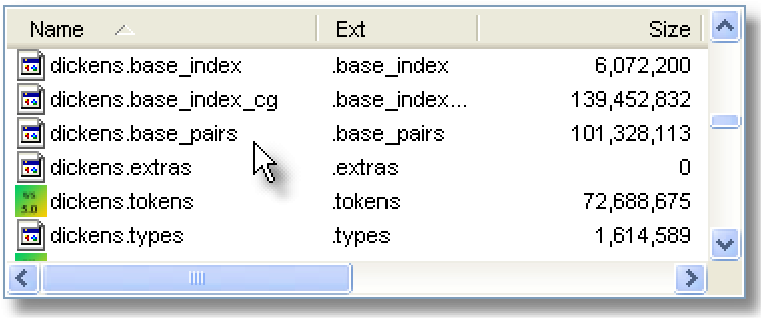
In the *Getting Started* window, first choose an existing Index, as here where an index based on the works of Dickens has been selected.



Getting Started ...

Index filename: F:\WSMITH\wordlist\32\index\dickens.tokens

To generate the concgrams, the program will then need to build some further files based on the existing index files:

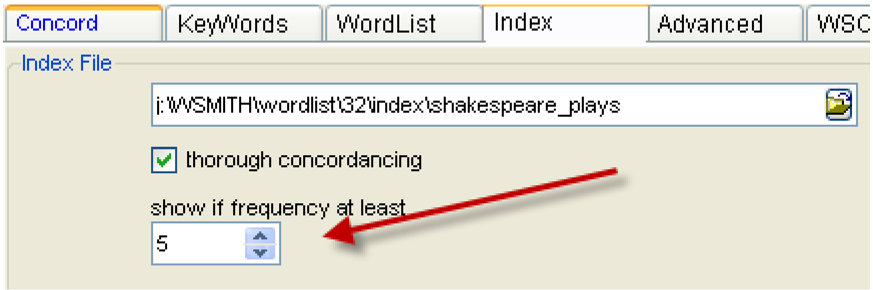


Name	Ext	Size
dickens.base_index	.base_index	6,072,200
dickens.base_index_cg	.base_index...	139,452,832
dickens.base_pairs	.base_pairs	101,328,113
dickens.extras	.extras	0
dickens.tokens	.tokens	72,688,675
dickens.types	.types	1,614,589

There are two steps simply because there's a lot of work if the original index is large. You can stop after the first stage and resume the next day if you wish. With a modern PC and a source text corpus of only a few million words, though, it should be possible to generate the files in a matter of a few minutes.

As you see above, some large additional files have been generated at the end of the two *Build steps* marked on the buttons in the top window.

All items which are found together at least as often as set in the Index settings (here 5 times)



Index File

i:\WSMITH\wordlist\32\index\shakespeare_plays

☑ thorough concordancing

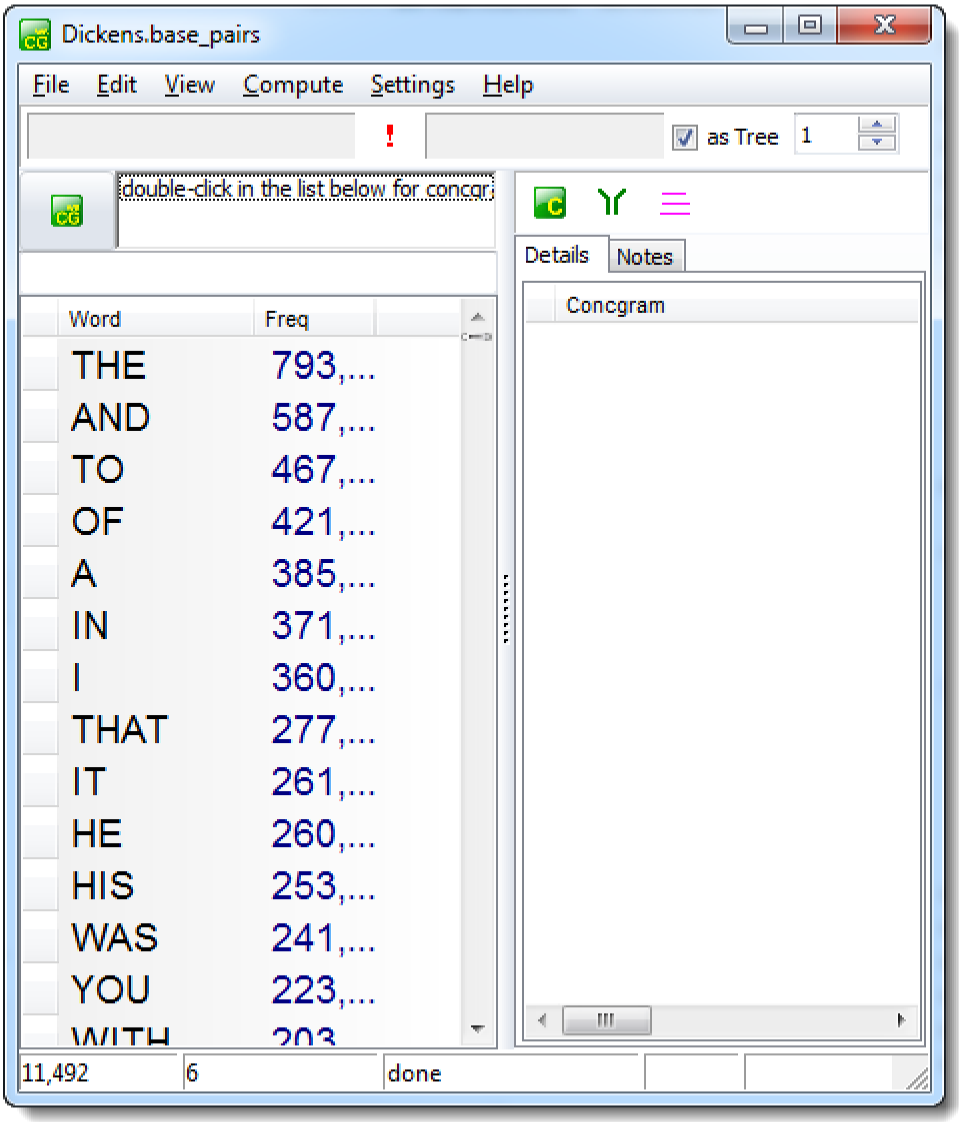
show if frequency at least 5

will be saved as potential members of each concgram.

Now, choose Show to view 397 the results. (Or, as usual, right-click the main WSConcgram window and choose *last file*).

10.9.5 viewing concgrams

When you first open a concgram file created by WSConcGram, it will look something like this one

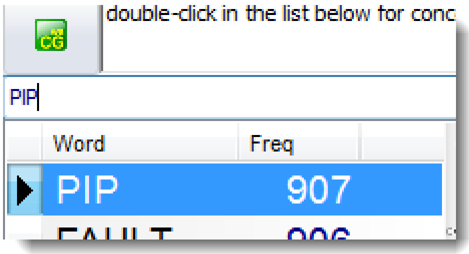


Dickens.base_pairs
File Edit View Compute Settings Help
☑ as Tree 1
double-click in the list below for concgr
Details Notes
Concgram

Word	Freq
THE	793,...
AND	587,...
TO	467,...
OF	421,...
A	385,...
IN	371,...
I	360,...
THAT	277,...
IT	261,...
HE	260,...
HIS	253,...
WAS	241,...
YOU	223,...
WITH	203

11,492 6 done

It'll appear (by default) in frequency order as set in the settings[395] but you can sort it by pressing the *Word* and *Freq* headers, and can search for items using the little box above the list.

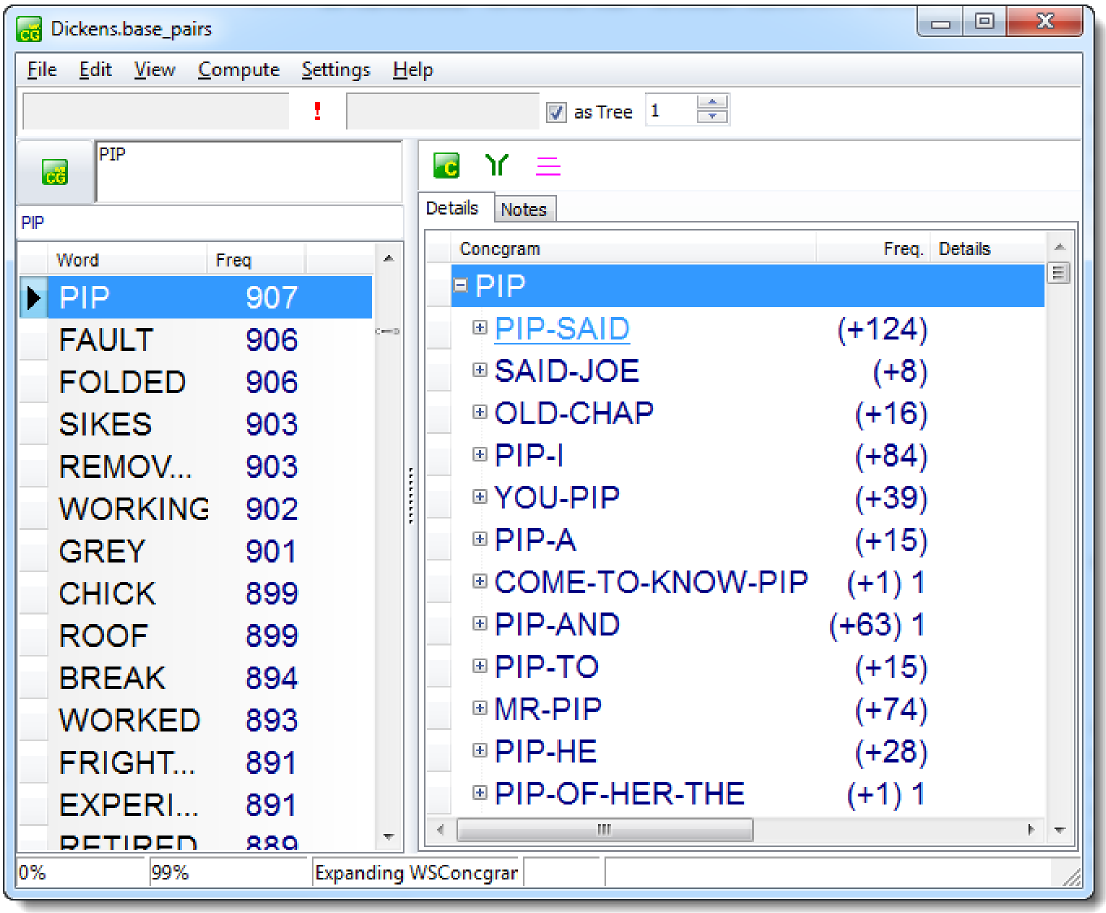


PIP

Word	Freq
PIP	907
FAULT	906

To get a detailed set of concgrams, double-click an item such as **PIP** (the hero of *Great Expectations*), or drag it to the list-box above. Then press the concgram button beside that.

You then get a tree view like this



☑ as Tree 1

PIP

Word	Freq
PIP	907
FAULT	906
FOLDED	906
SIKES	903
REMOV...	903
WORKING	902
GREY	901
CHICK	899
ROOF	899
BREAK	894
WORKED	893
FRIGHT...	891
EXPERI...	891
RETIRED	889

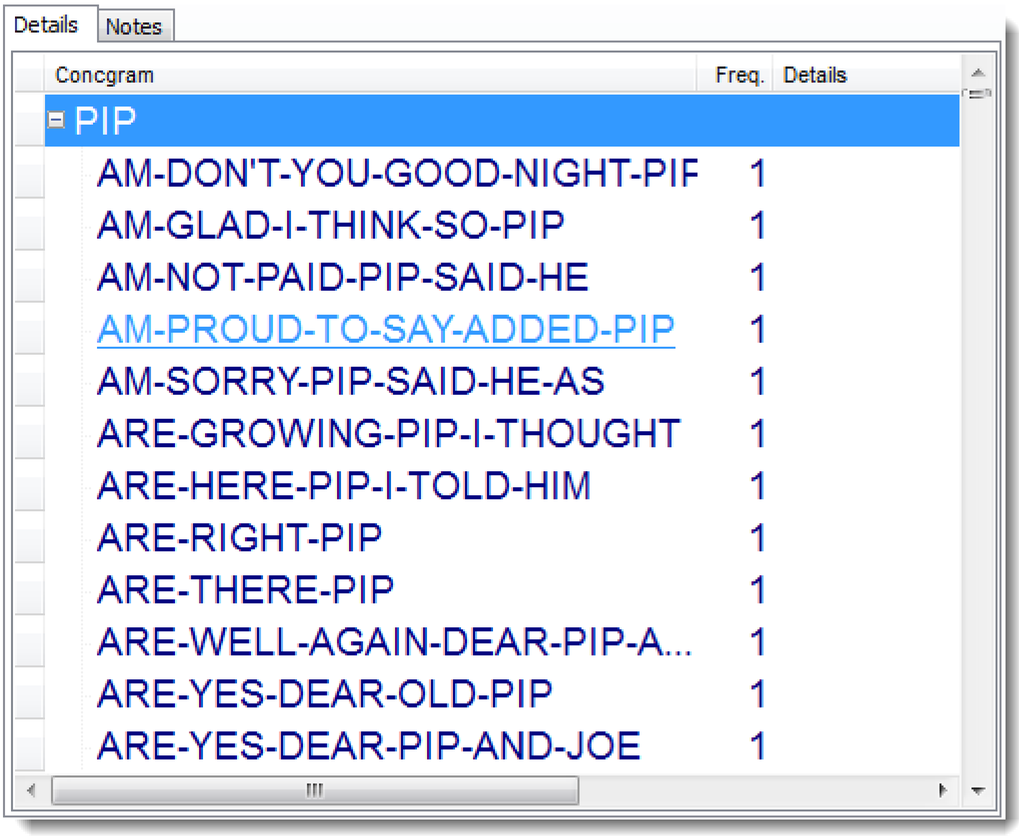
Concgram	Freq.	Details
⊟ PIP		
⊞ PIP-SAID	(+124)	
⊞ SAID-JOE	(+8)	
⊞ OLD-CHAP	(+16)	
⊞ PIP-I	(+84)	
⊞ YOU-PIP	(+39)	
⊞ PIP-A	(+15)	
⊞ COME-TO-KNOW-PIP	(+1) 1	
⊞ PIP-AND	(+63) 1	
⊞ PIP-TO	(+15)	
⊞ MR-PIP	(+74)	
⊞ PIP-HE	(+28)	
⊞ PIP-OF-HER-THE	(+1) 1	

0% 99% Expanding WSConcgrar

where similar items are grouped. Each branch of the tree shows how many sub-items and how many items of its own it has.

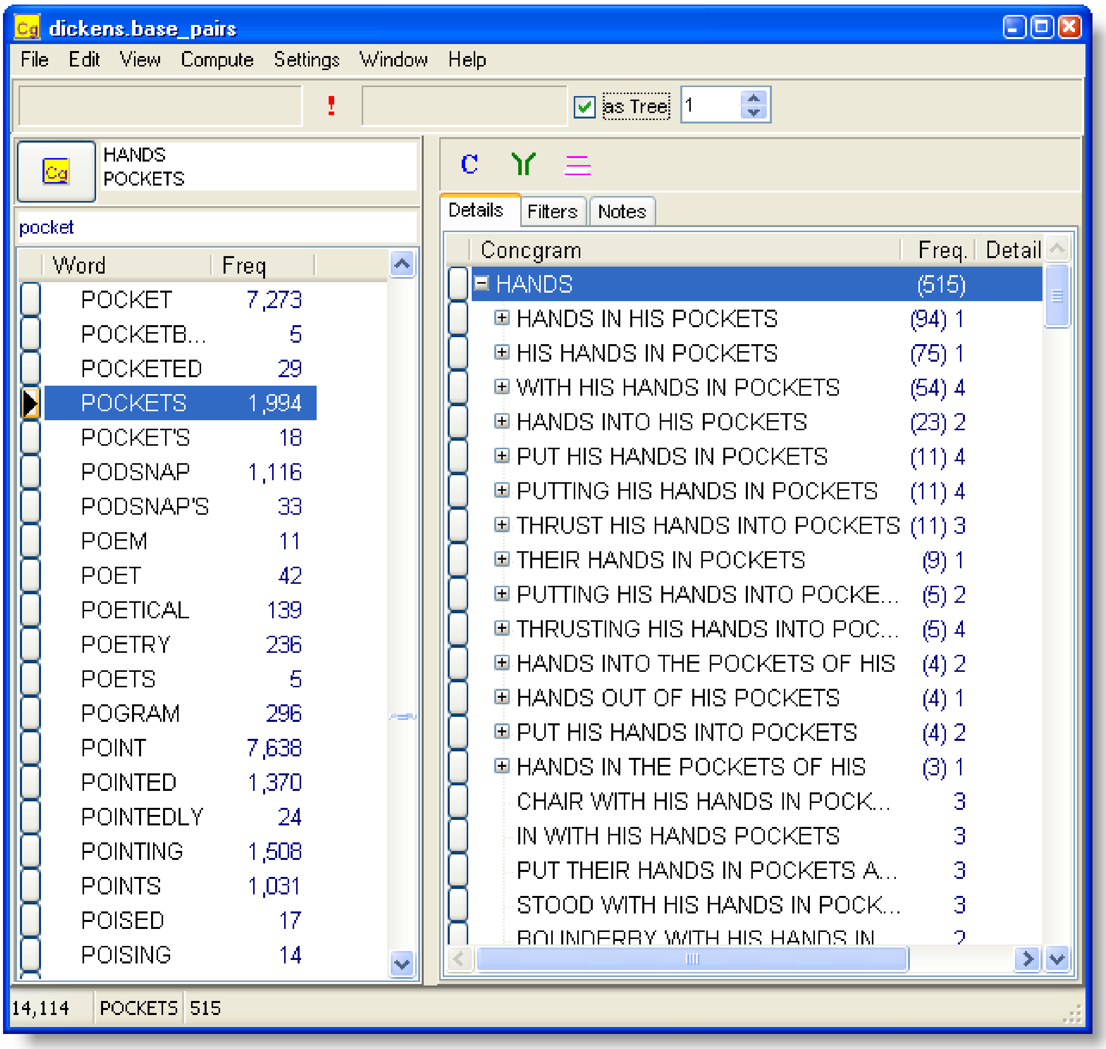
The other controls are used for suspending lengthy processing (!) changing from a tree-view to a list, for concordancing (C), for filtering 402 (Y), clearing filters (≡), and showing more or less of the tree (1).

So if you prefer a plain list, click *as Tree* to view like this:



Concgram	Freq.	Details
PIP		
AM-DON'T-YOU-GOOD-NIGHT-PIF	1	
AM-GLAD-I-THINK-SO-PIP	1	
AM-NOT-PAID-PIP-SAID-HE	1	
AM-PROUD-TO-SAY-ADDED-PIP	1	
AM-SORRY-PIP-SAID-HE-AS	1	
ARE-GROWING-PIP-I-THOUGHT	1	
ARE-HERE-PIP-I-TOLD-HIM	1	
ARE-RIGHT-PIP	1	
ARE-THERE-PIP	1	
ARE-WELL-AGAIN-DEAR-PIP-A...	1	
ARE-YES-DEAR-OLD-PIP	1	
ARE-YES-DEAR-PIP-AND-JOE	1	

You may if you like select several items like this:



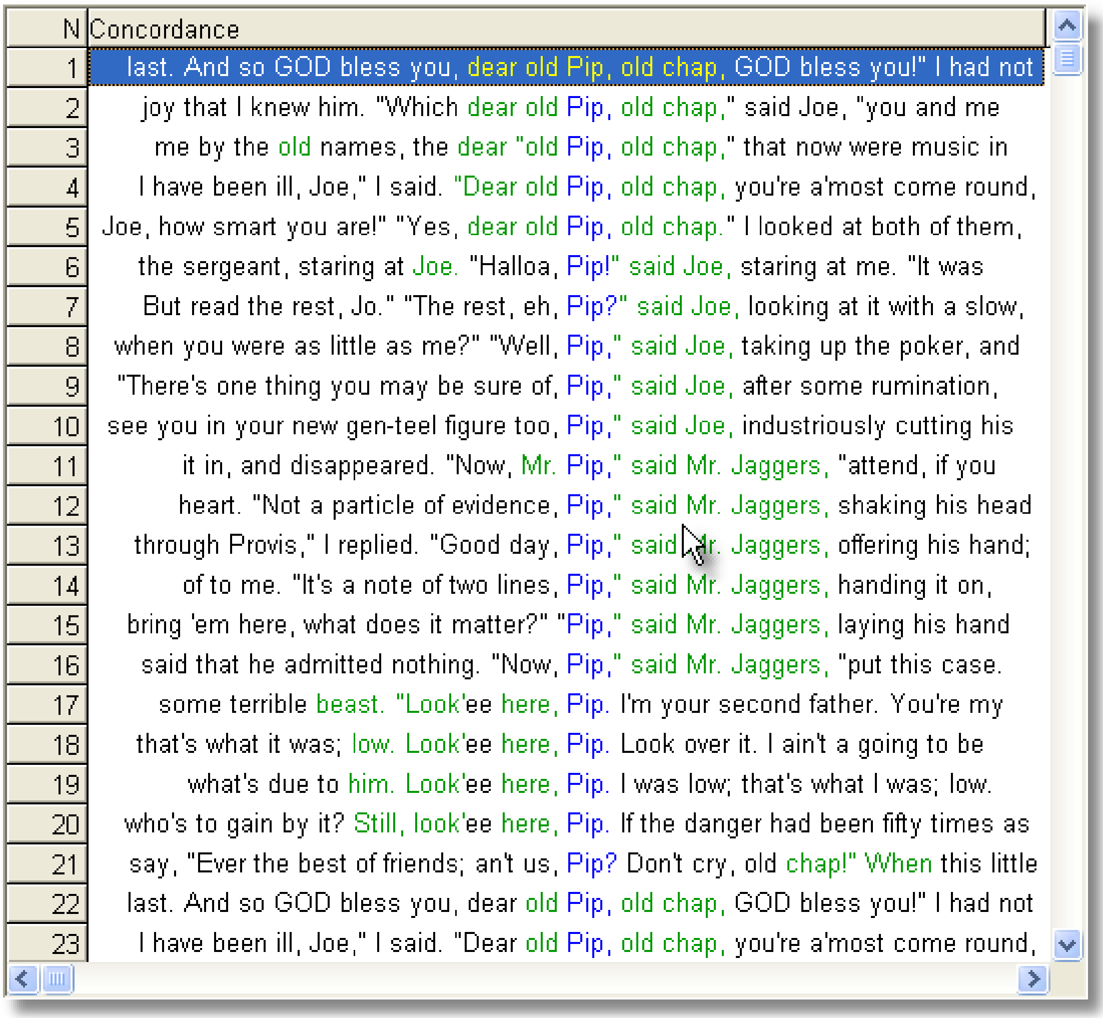
Word	Freq
POCKET	7,273
POCKETB...	5
POCKETED	29
POCKETS	1,994
POCKET'S	18
PODSNAP	1,116
PODSNAP'S	33
POEM	11
POET	42
POETICAL	139
POETRY	236
POETS	5
POGRAM	296
POINT	7,638
POINTED	1,370
POINTEDLY	24
POINTING	1,508
POINTS	1,031
POISED	17
POISING	14

Concgram	Freq.	Detail
⊟ HANDS	(515)	
⊞ HANDS IN HIS POCKETS	(94) 1	
⊞ HIS HANDS IN POCKETS	(75) 1	
⊞ WITH HIS HANDS IN POCKETS	(54) 4	
⊞ HANDS INTO HIS POCKETS	(23) 2	
⊞ PUT HIS HANDS IN POCKETS	(11) 4	
⊞ PUTTING HIS HANDS IN POCKETS	(11) 4	
⊞ THRUST HIS HANDS INTO POCKETS	(11) 3	
⊞ THEIR HANDS IN POCKETS	(9) 1	
⊞ PUTTING HIS HANDS INTO POCKE...	(5) 2	
⊞ THRUSTING HIS HANDS INTO POC...	(5) 4	
⊞ HANDS INTO THE POCKETS OF HIS	(4) 2	
⊞ HANDS OUT OF HIS POCKETS	(4) 1	
⊞ PUT HIS HANDS INTO POCKETS	(4) 2	
⊞ HANDS IN THE POCKETS OF HIS	(3) 1	
CHAIR WITH HIS HANDS IN POCK...	3	
IN WITH HIS HANDS POCKETS	3	
PUT THEIR HANDS IN POCKETS A...	3	
STOOD WITH HIS HANDS IN POCK...	3	
BOUNDERBY WITH HIS HANDS IN	2	

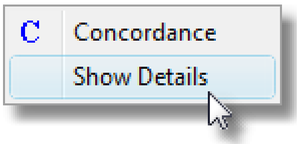
14,114 | POCKETS | 515

but do note that the concgrams will have to contain *all* of the words selected.

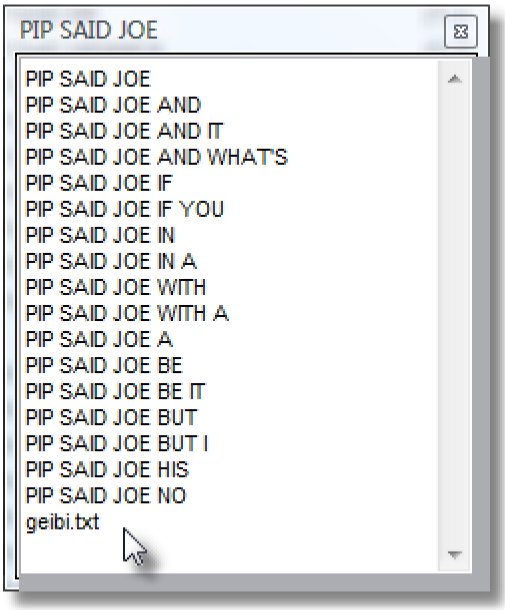
After filtering 402 appropriately and pressing the Concordance button



N	Concordance
1	last. And so GOD bless you, dear old Pip, old chap, GOD bless you!" I had not
2	joy that I knew him. "Which dear old Pip, old chap," said Joe, "you and me
3	me by the old names, the dear "old Pip, old chap," that now were music in
4	I have been ill, Joe," I said. "Dear old Pip, old chap, you're a'most come round,
5	Joe, how smart you are!" "Yes, dear old Pip, old chap." I looked at both of them,
6	the sergeant, staring at Joe. "Halloa, Pip!" said Joe, staring at me. "It was
7	But read the rest, Jo." "The rest, eh, Pip?" said Joe, looking at it with a slow,
8	when you were as little as me?" "Well, Pip," said Joe, taking up the poker, and
9	"There's one thing you may be sure of, Pip," said Joe, after some rumination,
10	see you in your new gen-teel figure too, Pip," said Joe, industriously cutting his
11	it in, and disappeared. "Now, Mr. Pip," said Mr. Jaggers, "attend, if you
12	heart. "Not a particle of evidence, Pip," said Mr. Jaggers, shaking his head
13	through Provis," I replied. "Good day, Pip," said Mr. Jaggers, offering his hand;
14	of to me. "It's a note of two lines, Pip," said Mr. Jaggers, handing it on,
15	bring 'em here, what does it matter?" "Pip," said Mr. Jaggers, laying his hand
16	said that he admitted nothing. "Now, Pip," said Mr. Jaggers, "put this case.
17	some terrible beast. "Look'ee here, Pip. I'm your second father. You're my
18	that's what it was; low. Look'ee here, Pip. Look over it. I ain't a going to be
19	what's due to him. Look'ee here, Pip. I was low; that's what I was; low.
20	who's to gain by it? Still, look'ee here, Pip. If the danger had been fifty times as
21	say, "Ever the best of friends; an't us, Pip? Don't cry, old chap!" When this little
22	last. And so GOD bless you, dear old Pip, old chap, GOD bless you!" I had not
23	I have been ill, Joe," I said. "Dear old Pip, old chap, you're a'most come round,



If you right-click and choose Show Details you'll get to see the details of any section of the tree you have selected:


PIP SAID JOE

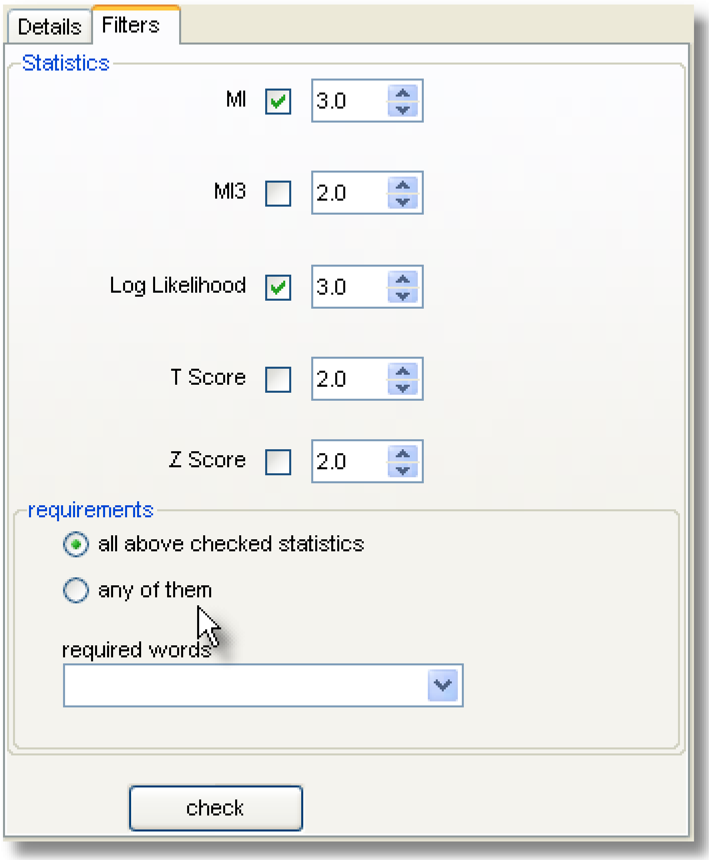
PIP SAID JOE
PIP SAID JOE AND
PIP SAID JOE AND IT
PIP SAID JOE AND WHAT'S
PIP SAID JOE IF
PIP SAID JOE IF YOU
PIP SAID JOE IN
PIP SAID JOE IN A
PIP SAID JOE WITH
PIP SAID JOE WITH A
PIP SAID JOE A
PIP SAID JOE BE
PIP SAID JOE BE IT
PIP SAID JOE BUT
PIP SAID JOE BUT I
PIP SAID JOE HIS
PIP SAID JOE NO
geibi.txt

where you see the various forms and the filename(s) they came from.

10.9.6 filtering concgrams

In order to select which items are "associated", we need some sort of suitable statistical procedures. The members of each concgram are at present merely associated by co-occurring at least a certain number of times as explained in generating 395 them

The Filtering settings in the Controller allow you to specify, for example, that you want to see only those which are associated with a MI (mutual information) score of 2.0 or a Log Likelihood score of 3.0.



MI ☑ 3.0

MI3 □ 2.0

Log Likelihood ☑ 3.0

T Score □ 2.0

Z Score □ 2.0

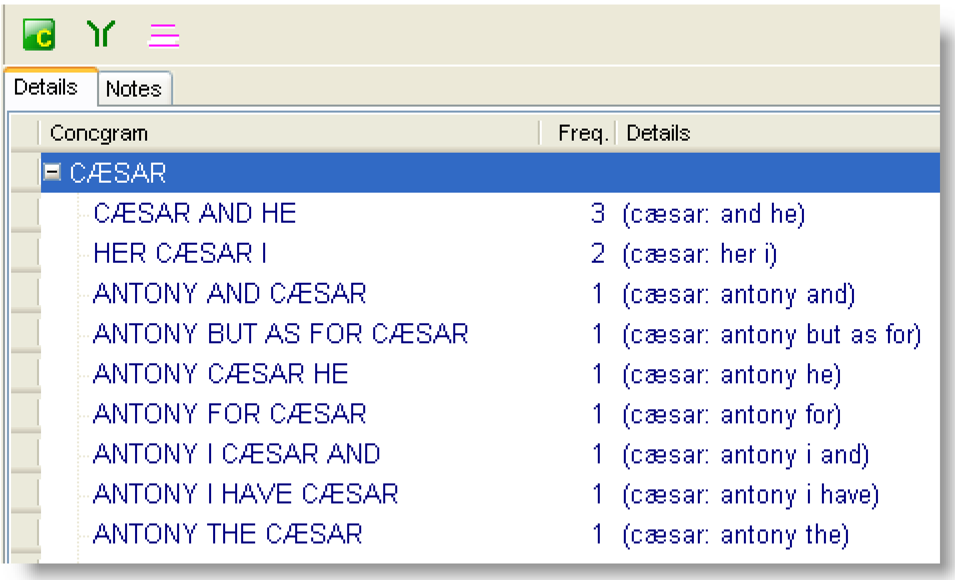
● all above checked statistics

○ any of them

required words

Ensure the statistics you need are checked and set to suitable thresholds, and decide whether all the thresholds have to be met (in the case above both MI and Log Likelihood would have to score 3.0 at least) or any of them (in the case above MI at 3.0 or above or Log Likelihood at 3.0 or above). You can also optionally insist on certain words being in your filtered results.

When you press the filter button (Y), you will see something like this:



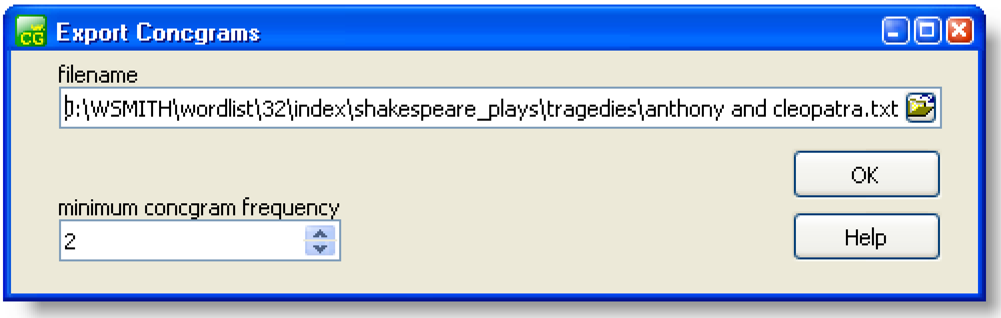
Concgram	Freq.	Details
⊟ CÆSAR		
CÆSAR AND HE	3	(cæsar: and he)
HER CÆSAR I	2	(cæsar: her i)
ANTONY AND CÆSAR	1	(cæsar: antony and)
ANTONY BUT AS FOR CÆSAR	1	(cæsar: antony but as for)
ANTONY CÆSAR HE	1	(cæsar: antony he)
ANTONY FOR CÆSAR	1	(cæsar: antony for)
ANTONY I CÆSAR AND	1	(cæsar: antony i and)
ANTONY I HAVE CÆSAR	1	(cæsar: antony i have)
ANTONY THE CÆSAR	1	(cæsar: antony the)

where the items which meet the filter requirements are separated out and selected ready for concordancing; any others are hidden. To the right you see that the head-word **CAESAR** here relates to **AND HE, HER, I , ANTONY** etc. above the thresholds set.

10.9.7 exporting concgrams

With concgram data loaded, you may wish to export it to a plain text file which can be imported into Excel or imported into a WordSmith word-list 307.

Choose *Compute* | *WordList* and you will be offered choices like these.



Export Concgrams

filename
D:\WSMITH\wordlist\32\index\shakespeare_plays\tragedies\anthony and cleopatra.txt

minimum concgram frequency
2

The suggested filename is based on your concgram data.

10.10 Character Profiler

10.10.1 purpose

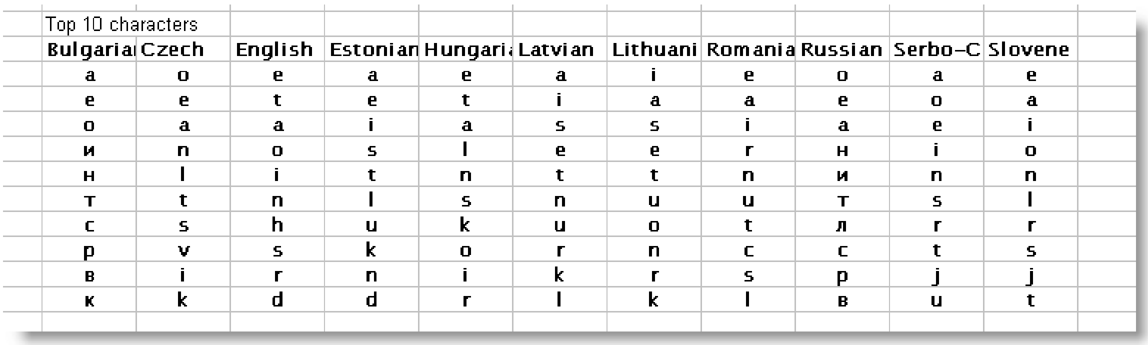
The point of it...

Character Profiler Ω, a tool to help find out which characters are most frequent in a text or a set of texts.

The purpose could be to check out which characters are most frequent (e.g. in normal English text the letter E followed by T will be most frequent as shown below), or it could be to check whether your text collection contains any oddities, such as accented characters or curly apostrophes you weren't expecting.

The first 32 codes used in computer storage of text are "control characters" such as tabs, line-feeds and carriage-returns. A plain **.txt** version of a text should only contain letters, numbers, punctuation and tabs, line-feeds and carriage-returns -- if there are other symbols you do not recognise you may have a **.txt** file which is really an old WordPerfect or Word **.doc** in disguise.

It would enable you to discover the most used characters across languages, as in this screenshot:



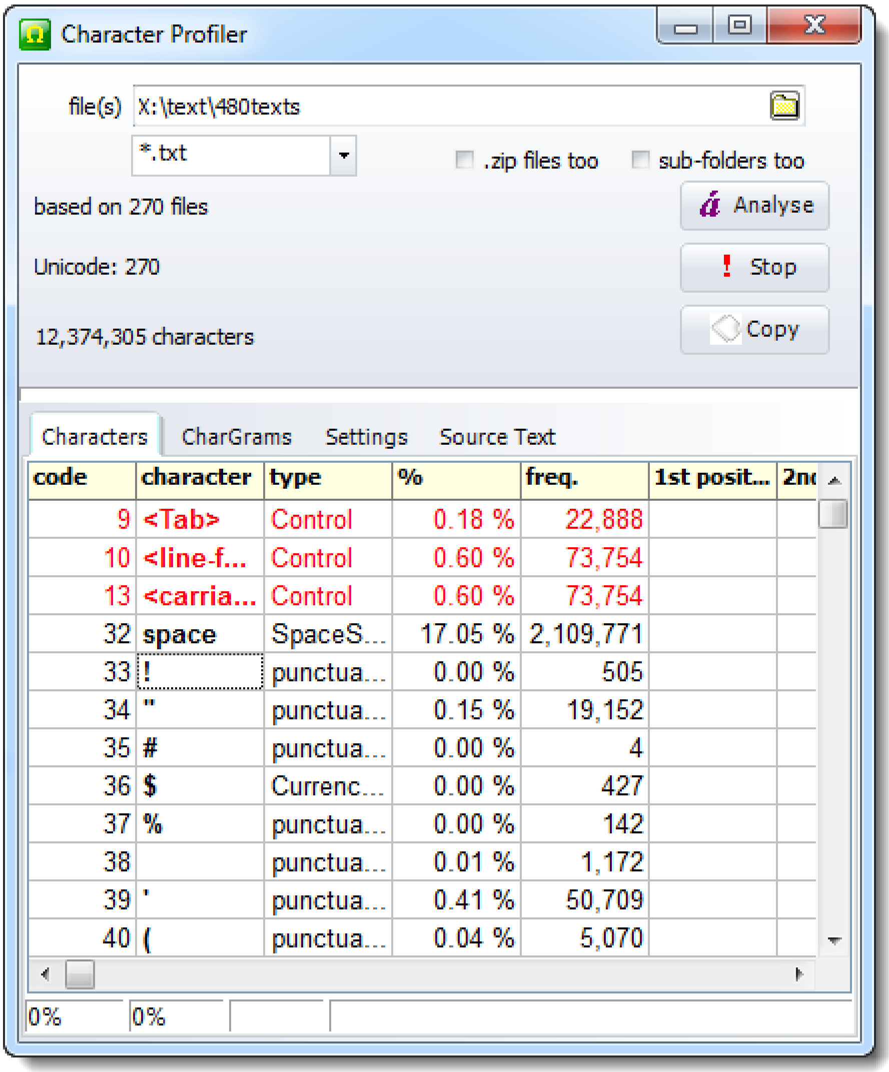
Top 10 characters										
Bulgaria	Czech	English	Estonian	Hungari	Latvian	Lithuani	Romania	Russian	Serbo-C	Slovene
а	o	e	a	e	a	i	e	о	a	e
е	e	t	e	t	i	a	a	е	o	a
о	a	a	i	a	s	s	i	а	e	i
и	n	o	s	l	e	e	r	н	i	o
н	l	i	t	n	t	t	n	и	n	n
т	t	n	l	s	n	u	u	т	s	l
с	s	h	u	k	u	o	t	л	r	r
р	v	s	k	o	r	n	c	с	t	s
в	i	r	n	i	k	r	s	р	j	j
к	k	d	d	r	l	k	l	в	u	t

For further details see http://lexically.net/downloads/corpus_linguistics/1984_characters.xls.

10.10.2 profiling characters

How to do it

1. Choose one or more texts or a folder. You can type in a complete filename (including drive and folder), and can use wildcards such as ***.txt**, or you can browse to find your text or folder.
2. If you want to study one text only, just choose one text, but you may choose a whole folderful or more by using the "sub-folders too" option.
3. Press *Analyse*.



file(s) X:\text\480texts

*.txt

□ .zip files too □ sub-folders too

based on 270 files

Unicode: 270

12,374,305 characters

code	character	type	%	freq.	1st posit...	2n(
9	<Tab>	Control	0.18 %	22,888		
10	<line-f...	Control	0.60 %	73,754		
13	<carria...	Control	0.60 %	73,754		
32	space	SpaceS...	17.05 %	2,109,771		
33	!	punctua...	0.00 %	505		
34	"	punctua...	0.15 %	19,152		
35	#	punctua...	0.00 %	4		
36	$	Currenc...	0.00 %	427		
37	%	punctua...	0.00 %	142		
38		punctua...	0.01 %	1,172		
39	'	punctua...	0.41 %	50,709		
40	(	punctua...	0.04 %	5,070		

0% 0%

The display shows details of your selected text, and if you click the *Source Text* tab you can see the original text. (If you have analysed a whole set of text files the *Source Text* tab will show only that last one.)

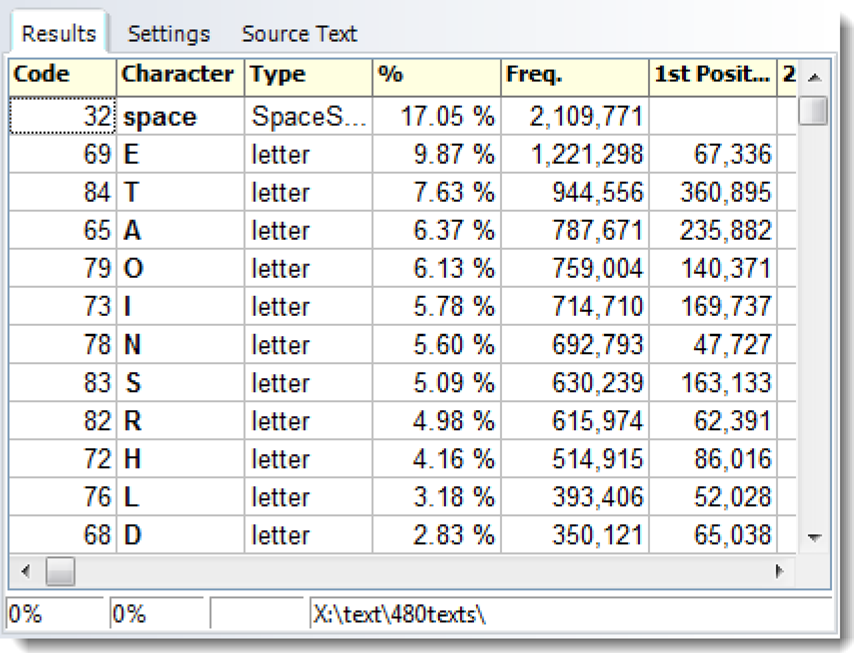
Legend

code	the Unicode code of
character	the character
type	distinguishing punctuation, digits, letters
%	percentage of the total number of characters in the text(s)
freq.	number of occurrences of that character
<Tab> etc.	control characters indicated in red.
1st Position	number of each letter-character occurring in word-initial position
2nd	number found in second position in any word
etc.	

Note that 8th will only be able to count letter frequencies for words at least 8 letters long, while 1st or 2nd will handle nearly all words.

Sort

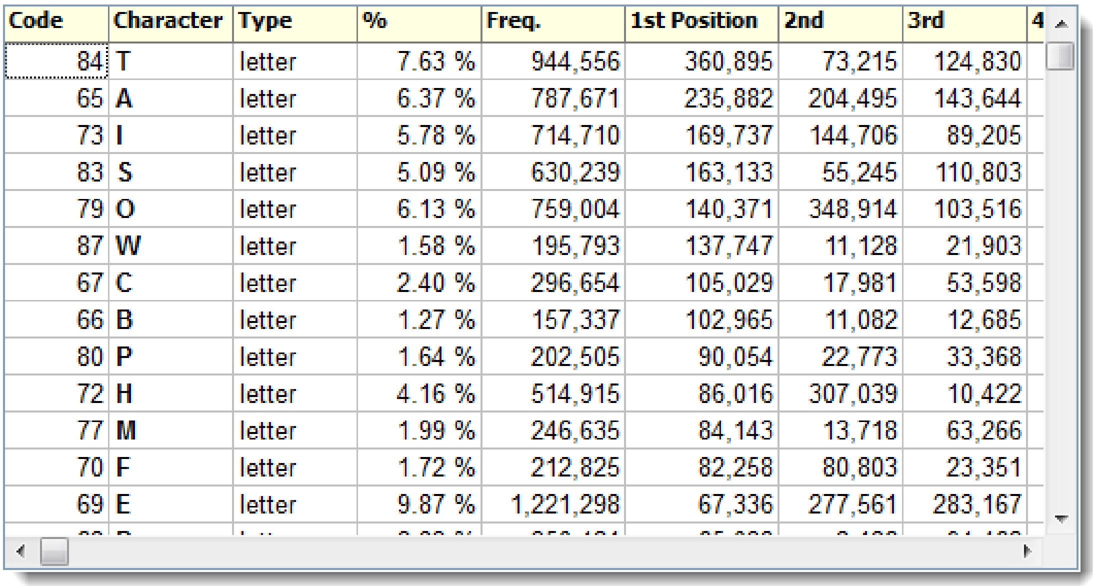
Click the header to sort the data:



Code	Character	Type	%	Freq.	1st Posit...	2
32	space	SpaceS...	17.05 %	2,109,771		
69	E	letter	9.87 %	1,221,298	67,336	
84	T	letter	7.63 %	944,556	360,895	
65	A	letter	6.37 %	787,671	235,882	
79	O	letter	6.13 %	759,004	140,371	
73	I	letter	5.78 %	714,710	169,737	
78	N	letter	5.60 %	692,793	47,727	
83	S	letter	5.09 %	630,239	163,133	
82	R	letter	4.98 %	615,974	62,391	
72	H	letter	4.16 %	514,915	86,016	
76	L	letter	3.18 %	393,406	52,028	
68	D	letter	2.83 %	350,121	65,038	

0% | 0% | | X:\text\480texts\

The letter **E** (upper and lower case merged) here represents nearly ten percent of all letters, closely followed by **T**. If sorted by 1st position in the word, however,



Code	Character	Type	%	Freq.	1st Position	2nd	3rd	4
84	T	letter	7.63 %	944,556	360,895	73,215	124,830	
65	A	letter	6.37 %	787,671	235,882	204,495	143,644	
73	I	letter	5.78 %	714,710	169,737	144,706	89,205	
83	S	letter	5.09 %	630,239	163,133	55,245	110,803	
79	O	letter	6.13 %	759,004	140,371	348,914	103,516	
87	W	letter	1.58 %	195,793	137,747	11,128	21,903	
67	C	letter	2.40 %	296,654	105,029	17,981	53,598	
66	B	letter	1.27 %	157,337	102,965	11,082	12,685	
80	P	letter	1.64 %	202,505	90,054	22,773	33,368	
72	H	letter	4.16 %	514,915	86,016	307,039	10,422	
77	M	letter	1.99 %	246,635	84,143	13,718	63,266	
70	F	letter	1.72 %	212,825	82,258	80,803	23,351	
69	E	letter	9.87 %	1,221,298	67,336	277,561	283,167	

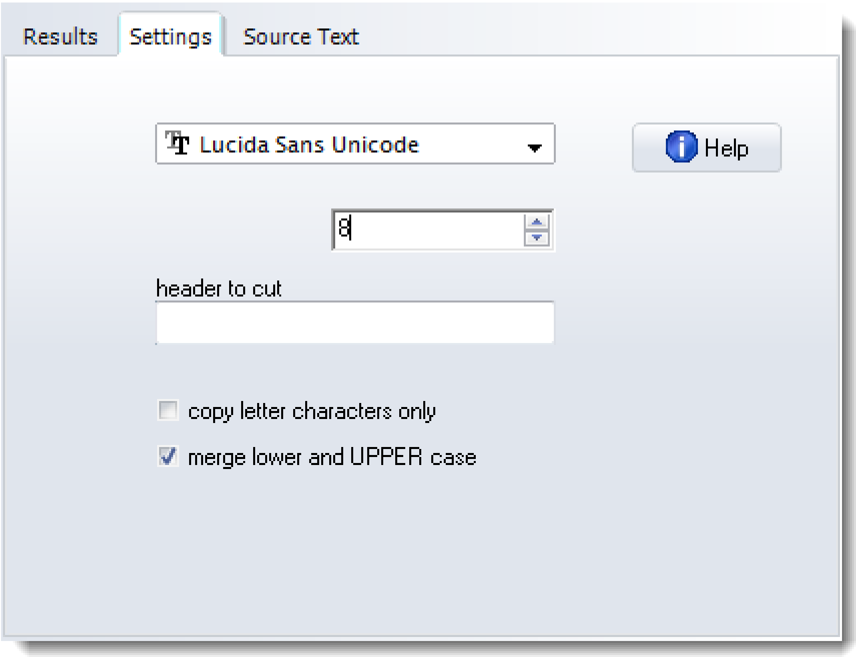
the letter **E** comes after **T, A, I, S, O, W, C, B, P, H, M** and **F** in frequency. Presumably the ranking of **T** reflects the frequency in English of **the** and **A** of **a**.

Copy

Copies the data to the clipboard, ready to be pasted for example into Excel.

See also: settings 408

10.10.3 profiling settings



Lucida Sans Unicode

8

header to cut

□ copy letter characters only

☑ merge lower and UPPER case

The top two boxes allow you to choose a font for your display. Most fonts can only represent some of the Unicode characters, so you may need to experiment to determine which is best for your language. (Character Profiler translates any text into Unicode whether or not it is in Unicode originally, and tells you which form it is in on the Results tab.)

Header to cut

If you've typed something in here such as **`</Header>`**, the program treats all the text before that as a header to be excluded from analysis..

Copy letter characters only

Check this one to force the copying to the clipboard to copy only data of letters, ignoring punctuation and digits.

Merge lower and UPPER case

Check this one to convert all text to upper case.

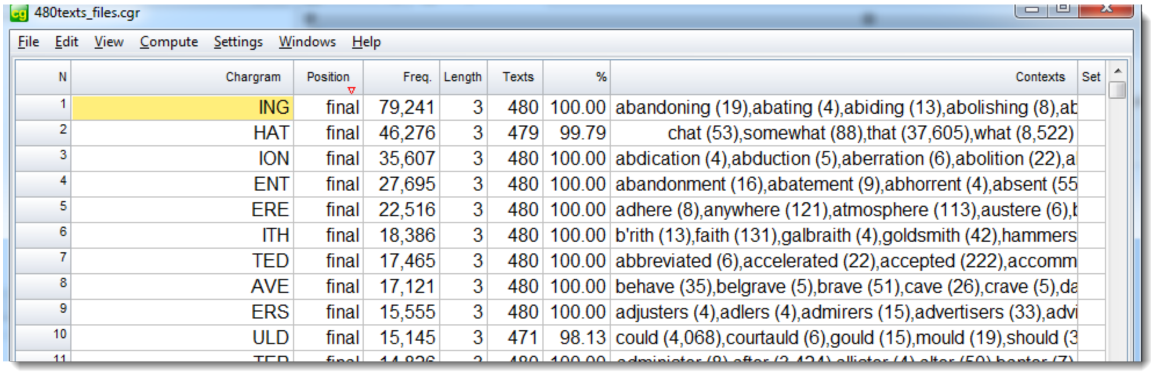
10.11 Chargrams

10.11.1 purpose

The point of it...

Chargrams, a tool to help find out which chargrams 426 (sequences of N characters) are most frequent in a text or a set of texts.

The purpose could be to check out which chargrams are most frequent e.g. in word-initial position, in the middle of a word, or at the end.



N	Chargram	Position	Freq.	Length	Texts	%	Contexts	Set
1	ING	final	79,241	3	480	100.00	abandoning (19),abating (4),abiding (13),abolishing (8),ab	
2	HAT	final	46,276	3	479	99.79	chat (53),somewhat (88),that (37,605),what (8,522)	
3	ION	final	35,607	3	480	100.00	abdication (4),abduction (5),aberration (6),abolition (22),a	
4	ENT	final	27,695	3	480	100.00	abandonment (16),abatement (9),abhorrent (4),absent (55	
5	ERE	final	22,516	3	480	100.00	adhere (8),anywhere (121),atmosphere (113),austere (6),t	
6	ITH	final	18,386	3	480	100.00	b'rith (13),faith (131),galbraith (4),goldsmith (42),hammers	
7	TED	final	17,465	3	480	100.00	abbreviated (6),accelerated (22),accepted (222),accomm	
8	AVE	final	17,121	3	480	100.00	behave (35),belgrave (5),brave (51),cave (26),crave (5),da	
9	ERS	final	15,555	3	480	100.00	adjusters (4),adlers (4),admirers (15),advertisers (33),advi	
10	ULD	final	15,145	3	471	98.13	could (4,068),courtauld (6),gould (15),mould (19),should (3	
11	TER	final	14,836	3	480	100.00	administer (8),after (2,424),alligator (4),alter (50),banter (7)	

These are 3-letter chargrams occurring in word-final position. **ING** is a well-known ending in English; **HAT** is is a frequent 3-letter sequence at the end of words too.

How does it work?

Chargrams are computed by taking only the valid characters of text. If a text contained "In 1845 there was a princess", the 3-character chargrams considered would be **THE, HER, ERE, WAS, PRI, RIN, INC, NCE, CES, ESS**. The positions are computed in relation to the original words, so **THE** is word-initial while **ESS** is word-final, and **RIN** is medial.

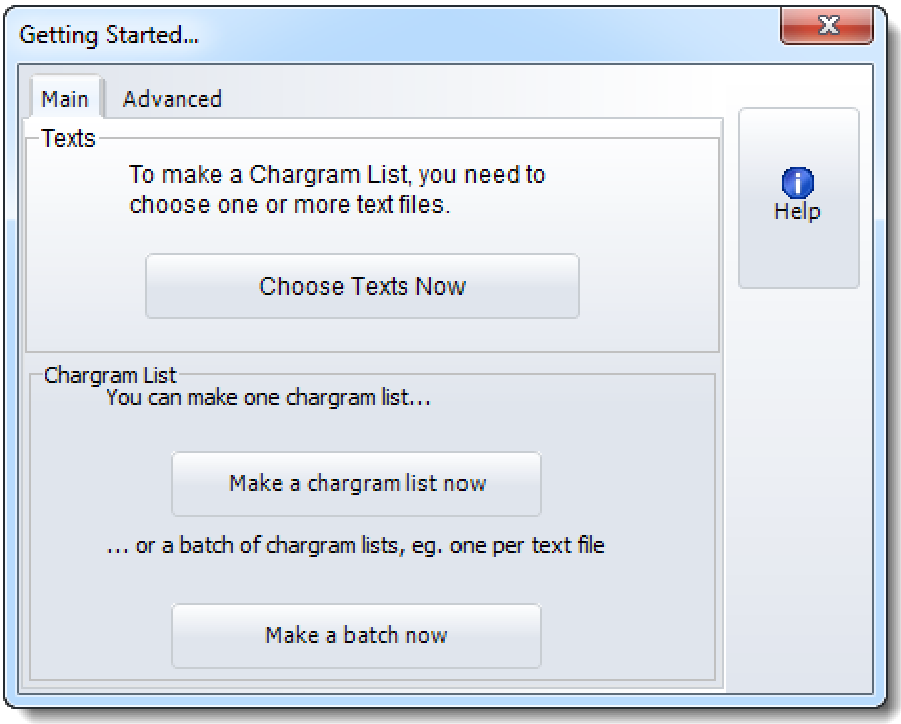
If including punctuation, the sequences would include **IN_, N_1, _18, 184** etc. too.

10.11.2 chargram procedure

How to do it

Choose *File* | *New* in the Chargrams menu.

Choose your texts as with the other Tools.



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

You can make one chargram list...

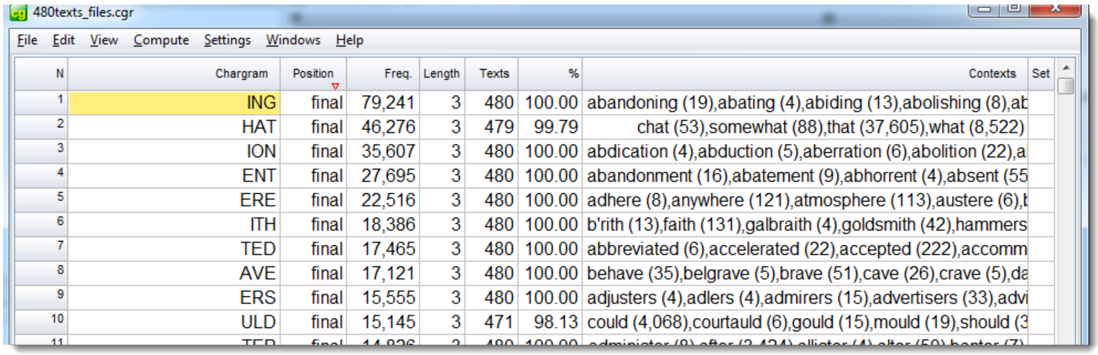
... or a batch of chargram lists, eg. one per text file

Then press the button to make a chargram list.

See also : settings 413.

10.11.3 display

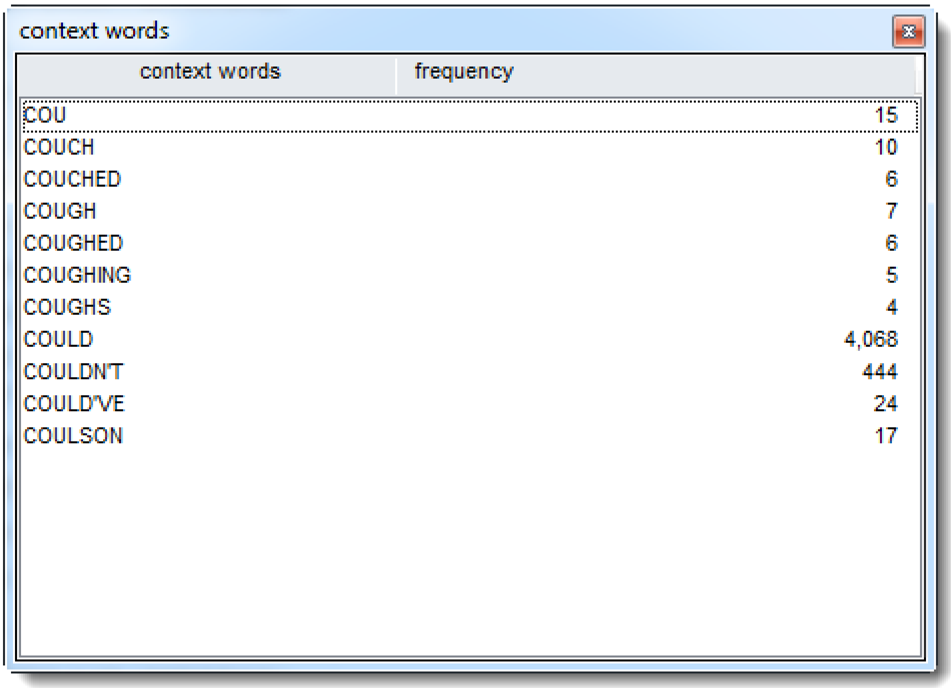
The display is similar to that in the WordList tool.



N	Chargram	Position	Freq.	Length	Texts	%	Contexts	Set
1	ING	final	79,241	3	480	100.00	abandoning (19),abating (4),abiding (13),abolishing (8),ab	
2	HAT	final	46,276	3	479	99.79	chat (53),somewhat (88),that (37,605),what (8,522)	
3	ION	final	35,607	3	480	100.00	abdication (4),abduction (5),aberration (6),abolition (22),a	
4	ENT	final	27,695	3	480	100.00	abandonment (16),abatement (9),abhorrent (4),absent (55	
5	ERE	final	22,516	3	480	100.00	adhere (8),anywhere (121),atmosphere (113),austere (6),b	
6	ITH	final	18,386	3	480	100.00	b'rith (13),faith (131),galbraith (4),goldsmith (42),hammers	
7	TED	final	17,465	3	480	100.00	abbreviated (6),accelerated (22),accepted (222),accomm	
8	AVE	final	17,121	3	480	100.00	behave (35),belgrave (5),brave (51),cave (26),crave (5),da	
9	ERS	final	15,555	3	480	100.00	adjusters (4),adlers (4),admirers (15),advertisers (33),advi	
10	ULD	final	15,145	3	471	98.13	could (4,068),courtauld (6),gould (15),mould (19),should (3	
11	TER	final	14,826	3	480	100.00	administer (8),after (2,424),alligator (4),alter (50),banter (7),	

Contexts

This column shows the word-contexts for each chargram. You can double-click to see the whole list.

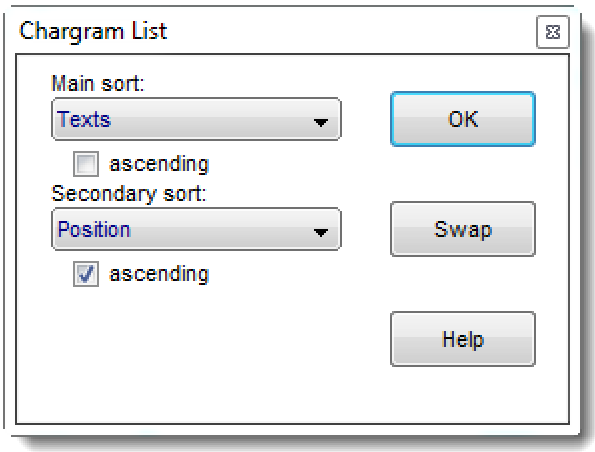


context words

context words	frequency
COU	15
COUCH	10
COUCHED	6
COUGH	7
COUGHED	6
COUGHING	5
COUGHS	4
COULD	4,068
COULDN'T	444
COULD'VE	24
COULSON	17

Sorting

You can sort by clicking a header. This offers you two of the columns to sort on, a primary sort and then where values on the primary sort are the same a secondary sort. In this example the user is choosing the number of texts in descending order and then word-position in ascending order.



Chargram List

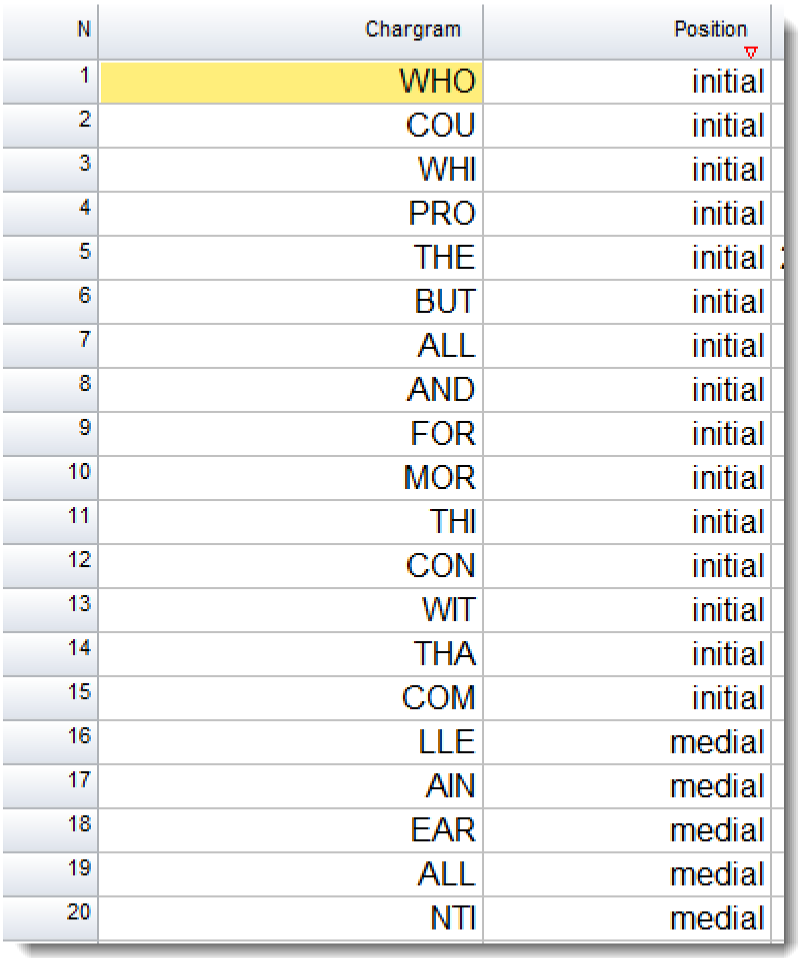
Main sort: Texts

□ ascending

Secondary sort: Position

☑ ascending

This gave the following list (extract):



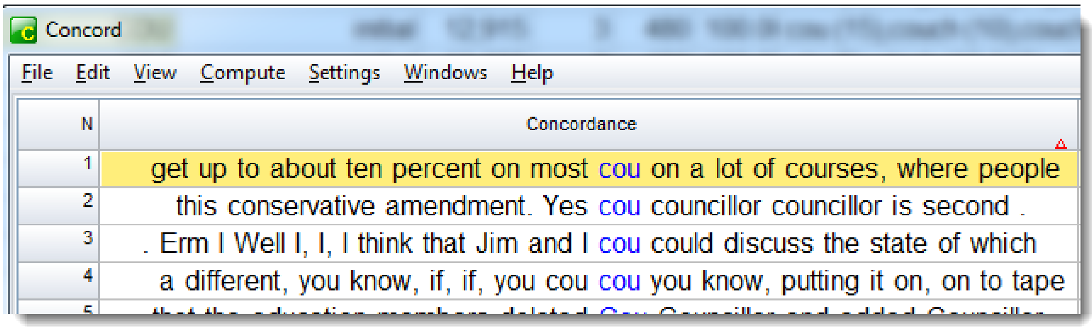
N	Chargram	Position
1	WHO	initial
2	COU	initial
3	WHI	initial
4	PRO	initial
5	THE	initial
6	BUT	initial
7	ALL	initial
8	AND	initial
9	FOR	initial
10	MOR	initial
11	THI	initial
12	CON	initial
13	WIT	initial
14	THA	initial
15	COM	initial
16	LLE	medial
17	AIN	medial
18	EAR	medial
19	ALL	medial
20	NTI	medial

Word-initial chargrams 1-15 and mid-word chargrams 16 onwards all occurred in all 100% of the texts selected.

Concordancing

As in many other Tools, you can concordance selected items by choosing *Compute* | *Concordance* in the menu.

In the case above I wondered about the context word **cou**



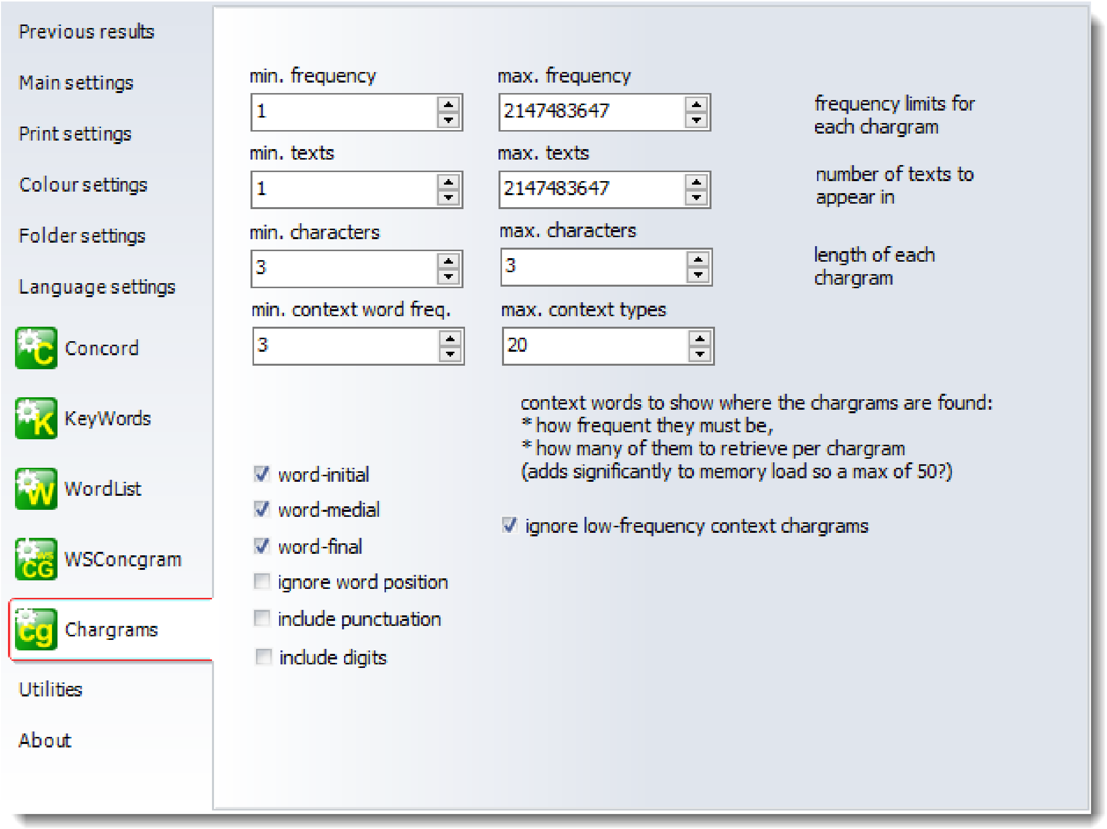
N	Concordance
1	get up to about ten percent on most cou on a lot of courses, where people
2	this conservative amendment. Yes cou councillor councillor is second .
3	. Erm I Well I, I, I think that Jim and I cou could discuss the state of which
4	a different, you know, if, if, you cou cou you know, putting it on, on to tape
5	that the education members deleted Cou Councillor and added Councillor

... which clearly shows speech reformulation.

See also : settings 413.

10.11.4 settings

Settings are found in the main Controller.



min. frequency: 1 | max. frequency: 2147483647 | frequency limits for each chargram

min. texts: 1 | max. texts: 2147483647 | number of texts to appear in

min. characters: 3 | max. characters: 3 | length of each chargram

min. context word freq.: 3 | max. context types: 20

context words to show where the chargrams are found:
* how frequent they must be,
* how many of them to retrieve per chargram
(adds significantly to memory load so a max of 50?)

☑ word-initial
☑ word-medial
☑ word-final
□ ignore word position
□ include punctuation
□ include digits

☑ ignore low-frequency context chargrams

You can set minimum and maximum token frequencies for the chargrams to be included in the results, a minimum and maximum number of texts they must appear in, the length in characters (e.g. 3 to 4 characters).

Context words

As the chargrams are selected, note is taken of the word which they are found in. Here you can determine a minimum number of times for each chargram to appear in a given word for that word to be listed, and a maximum number of context words per chargram to be collected. (Storing lots of extra words will use up system memory so a default of 20 or 50 may be reasonable.

Word position

By default chargrams in all three positions (word-initial, word-medial and word-final) will be collected. If you check the *ignore word position* box, word positions get merged.

Include punctuation

Allows chargrams of all characters (symbols, punctuation etc.) to be included. Spaces get replaced by underscores.

Include digits

Allows chargrams of digits as well as alphabetic characters.

Ignore low-frequency context chargrams

This setting allows us to filter out any chargrams which do not occur in many contexts. As shown here, any chargrams not occurring in at least 4 context word types will get eliminated. If, for example, a chargram has been found in 5 context word-types then the chargram is included in your list. But if in only 1 of these it is found occurring at least 4 times (i.e. found in the same context word-type recurring in the texts at least 4 times), you will see one context word only in the *contexts* column

See also: chargrams display 410

Reference

Section XI

11 Reference

11.1 acknowledgements

WordSmith Tools has developed over a period of years. Originally each tool came about because I wanted a tool for a particular job in my work as an Applied Linguist. Early versions were written for DOS, then Windows™ came onto the scene.

One tool, **Concord**, had a slightly different history. It developed out of *MicroConcord* which Tim Johns and I wrote for DOS and which Oxford University Press published in 1993.

The first published version was written in Borland™ Pascal with the time-critical sections in Assembler. Subsequently the programs were converted to Delphi™ 16-bit; this is a 32-bit only version written in Delphi XE and still using time-critical sections in Assembler.

I am grateful to

- lots of users who have made suggestions and given bug reports, for their feedback on aspects of the suite (including bugs!), and suggestions as to features it should have.
- generations of students and colleagues at the School of English, University of Liverpool, the MA Programme in Applied Linguistics at the Catholic University of São Paulo, colleagues and students at Aston University.
- Audrey Spina, Élodie Guthmann and Julia Hotter for their help with the French & German versions of WS 4.0; Spela Vintar's student for Slovenian; Zhu Yi and others at SFLEP in Shanghai for Mandarin for WS 5.0.
- Robert Jedrzejczyk (http://prog.olsztyn.pl/paslibvlc) for his PasLibVCLPlayer which enables WordSmith to play video.

Researchers from many other countries have also acted as alpha-testers and beta-testers and I thank them for their patience and feedback. I am also grateful to Nell Scott and other members of my family who have always given valuable support, feedback and suggestions.

Mike Scott
Feel free to email me at my contact address 425 with any further ideas for developing **WordSmith Tools**.

11.2 API

It is possible to run the WordSmith routines from your own programs; for this there's an API published. If you know a programming language, you can call a `.dll` which comes with WordSmith and ask it to create a concordance, a word-list or a key words list, which you can then process to suit your own purposes.

Easier, however, is to write a very simple batch script 31 which will run WordSmith unattended.

See also : custom processing 67

11.3 bibliography

Aston, Guy, 1995, "Corpora in Language Pedagogy: matching theory and practice", in G. Cook & B. Seidlhofer (eds.) *Principle & Practice in Applied Linguistics: Studies in honour of H.G. Widdowson*, Oxford: Oxford University Press, 257-70.

Aston, Guy & Burnard, Lou, 1998, *The BNC Handbook*, Edinburgh: Edinburgh University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan, 2000, *Longman Grammar of Spoken and Written English*, Harlow: Addison Wesley Longman.

Clear, Jeremy, 1993, "From Firth Principles: computational tools for the study of collocation" in M. Baker, G. Francis & E. Tognini-Bonelli (eds.), 1993, *Text and Technology: in honour of John Sinclair*, Philadelphia: John Benjamins, 271-92.

Cheng, Winnie, Chris Greaves & Martin Warren, 2006, From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, Vol .11, No. 4, pp. 411-433.

Dunning, Ted, 1993, "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics*, Vol 19, No. 1, pp. 61-74.

Fillmore, Charles J, & Atkins, B.T.S, 1994, "Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography", in B.T.S. Atkins & A. Zampolli, *Computational Approaches to the Lexicon*, Oxford:Clarendon Press, pp. 349-96.

Katz, Slava, 1996, Distribution of Common Words and Phrases in Text and Language Modelling, *Natural Language Engineering* 2 (1), 15-59

Murison-Bowie, Simon, 1993, *MicroConcord Manual: an introduction to the practices and principles of concordancing in language teaching*, Oxford: Oxford University Press.

Nakamura, Junsaku, 1993, "Statistical Methods and Large Corpora: a new tool for describing text types" in M. Baker, G. Francis & E. Tognini-Bonelli (eds.), 1993, *Text and Technology: in honour of John Sinclair*, Philadelphia: John Benjamins, 293-312.

Oakes, Michael P. 1998, *Statistics for Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Scott, Mike, 1997, "PC Analysis of Key Words - and Key Key Words", *System*, Vol. 25, No. 2, pp. 233-45.

Scott, Mike & Chris Tribble, 2006, *Textual Patterns: keyword and corpus analysis in language education*, Amsterdam: Benjamins.

Sinclair, John M, 1991, *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Stubbs, Michael, 1986, "Lexical Density: A Technique and Some Findings", in M. Coulthard (ed.) Talking About Text: Studies presented to David Brazil on his retirement, *Discourse Analysis Monograph no. 13*, Birmingham: English Language Research, Univ. of Birmingham, 27-42.

Stubbs, Michael, 1995, "Corpus Evidence for Norms of Lexical Collocation", in G. Cook & B. Seidlhofer (eds.) *Principle & Practice in Applied Linguistics: Studies in honour of H.G. Widdowson*, Oxford: Oxford University Press, 245-56.

Tuldava, J. 1995, *Methods in Quantitative Linguistics*, Trier: WVT Wissenschaftlicher Verlag Trier.

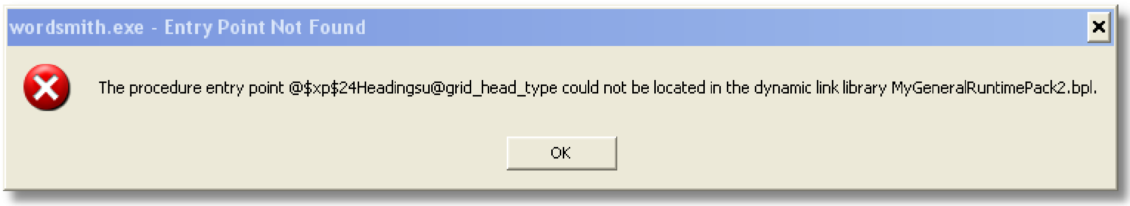
Youlmans, Gilbert, 1991, "A New Tool for Discourse Analysis: the vocabulary-management profile", *Language*, V. 67, No. 4, pp. 763-89.

UCREL's log likelihood information

11.4 bugs

All computer programs contain bugs. You may have seen a "General Protection Fault" message when using big expensive drawing or word-processing packages.

If you see something like this,



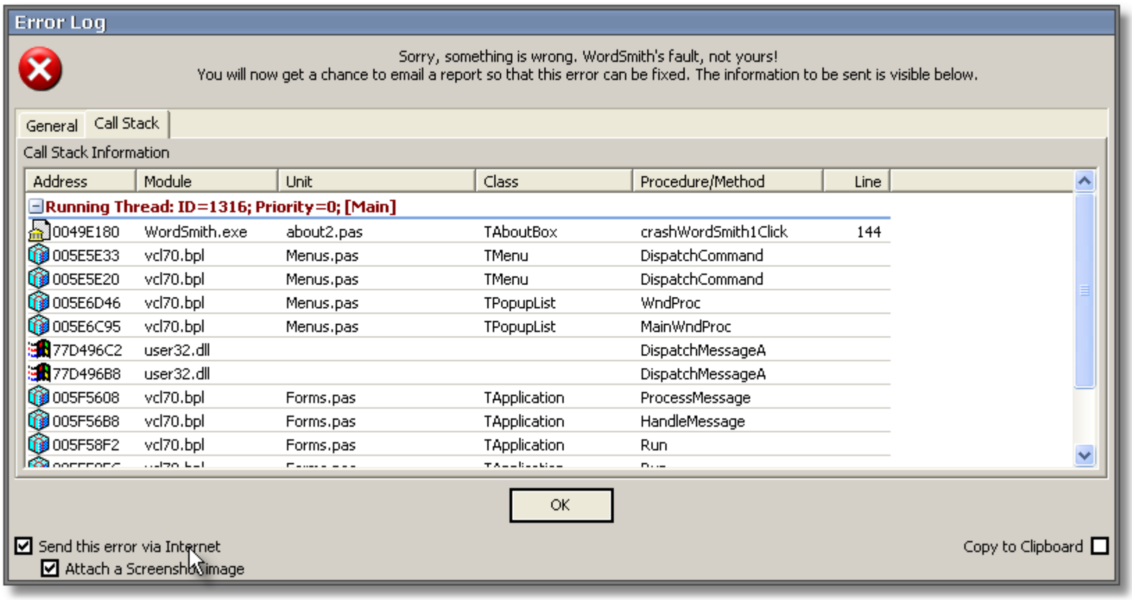
wordsmith.exe - Entry Point Not Found

The procedure entry point @xp24Headingsu@grid_head_type could not be located in the dynamic link library MyGeneralRuntimePack2.bpl.

then you have an incompatibility between sections of WordSmith. You have probably downloaded a fresh version of some parts of WordSmith but not all, and the various sub-programs are in conflict... The solution is a fresh download. http://lexically.net/wordsmith/version6/faqs/updating_or_reinstalling.htm explains.

Otherwise you should get a report popping up, giving "General" information about your PC and "Details" about the fault. This information will help me to fix the problem and will be saved in a small text file called **`wordsmith.elf`**, **`concord.elf`**, **`wordlist.elf`**, etc. When you quit the program, you will be offered a chance to email this to me.

The first thing you'll see when one of these happens is something like this:



Error Log

Sorry, something is wrong. WordSmith's fault, not yours!
You will now get a chance to email a report so that this error can be fixed. The information to be sent is visible below.

Call Stack Information

Address	Module	Unit	Class	Procedure/Method	Line
Running Thread: ID=1316; Priority=0; [Main]					
0049E180	WordSmith.exe	about2.pas	TAboutBox	crashWordSmith1Click	144
005E5E33	vcl70.bpl	Menus.pas	TMenu	DispatchCommand	
005E5E20	vcl70.bpl	Menus.pas	TMenu	DispatchCommand	
005E6D46	vcl70.bpl	Menus.pas	TPopupList	WndProc	
005E6C95	vcl70.bpl	Menus.pas	TPopupList	MainWndProc	
77D496C2	user32.dll			DispatchMessageA	
77D496B8	user32.dll			DispatchMessageA	
005F5608	vcl70.bpl	Forms.pas	TApplication	ProcessMessage	
005F56B8	vcl70.bpl	Forms.pas	TApplication	HandleMessage	
005F58F2	vcl70.bpl	Forms.pas	TApplication	Run	

☑ Send this error via Internet
☑ Attach a Screenshot image

Copy to Clipboard □

You may have to quit when you have pressed OK, or WordSmith may be able to cope despite the problem.
Usually the offending program will be able to cope despite the bug or you can go straight back into it without even needing to quit the main WordSmith Tools Controller 4, retrieve your saved results 101 from disk, and resume. If that doesn't work, try quitting WordSmith Tools overall, or quit Windows and then start it up again.

When you press OK, your email program should have a message with a couple of attachments to send to me.

The email message will only get sent when you press Send in your email program. It is only sent to me and I will not pass it on to anyone else. Read it first if you are worried about revealing your innermost secrets ... it will tell me the operating system, the amount of RAM and hard disk space, the version of WordSmith, and some technical details of routines

which it was going through when the crash occurred.

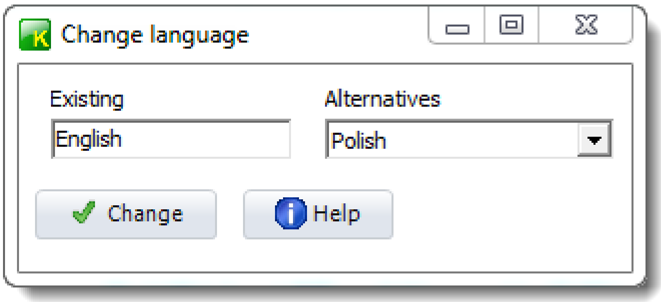
error messages 461

These warn you about problems which occur as the program works, e.g. if there's no room left on your disk, or you type in an impossible file name or a number containing a comma.

See also: logging 31, troubleshooting 455.

11.5 change language

If you have results computed with the wrong language setting, that can affect things, e.g. a key word listing depends on finding the words in the right order 315. To redefine the language of your data, choose *Edit* | *Change Language*, and in the resulting window



Existing: English

Alternatives: Polish

press *Change* once you have chosen a suitable alternative. If you choose a different one from the list of Alternatives, your Language and Text settings 124 in the main Controller will change too. In this screenshot, pressing *Change* will change the language to Polish.

11.6 Character Sets

11.6.1 overview

You need "plain text" in WordSmith. Not Microsoft Word .doc 444 files -- which contain text and a whole lot of other things too that you cannot normally see.

If you are processing English only, your texts can be in ASCII, ANSI or Unicode; WordSmith handles both formats. If in other languages, read on...

To handle a text in a computer, programs need to know how the text is encoded. In its processing, the software sees only a long string of numbers, and these have to match up with what you and I can recognise as "characters". For many languages like English with a restricted alphabet, encoding can be managed with only 1 "byte" per character. On the other hand a language like Chinese, which draws upon a very large array of characters, cannot easily be fitted to a 1-byte system. Hence the creation of other "multi-byte" systems. Obviously if a text in English is encoded in a multi-byte way, it will make a bigger file than one encoded with 1 byte per character, and this is slightly wasteful of disk and memory space. So, at the time of writing, 1-byte character sets are still in very widespread use. UTF-8 is a name for a multi-byte method, widely used for Chinese, etc.

In practice, your texts are likely to be encoded in a Windows 1-byte system, older texts in a DOS 1-byte system, and newer ones, especially in Chinese, Japanese, Greek, in Unicode. What matters most to you is what each character looks like, but WordSmith cannot possibly sort words correctly, or even recognise where a word begins and ends, if the encoding is not correct. WordSmith has to know (or try to find out) which system your texts are encoded in. It can perform certain tests in the background. But as it doesn't actually understand the words it sees, it is much safer for you to convert to Unicode, especially if you process texts in German, Spanish, Russian, Greek, Polish, Japanese, Farsi, Arabic etc.

Three main kinds of character set, each with its own flavours, are Windows, DOS, and Unicode.

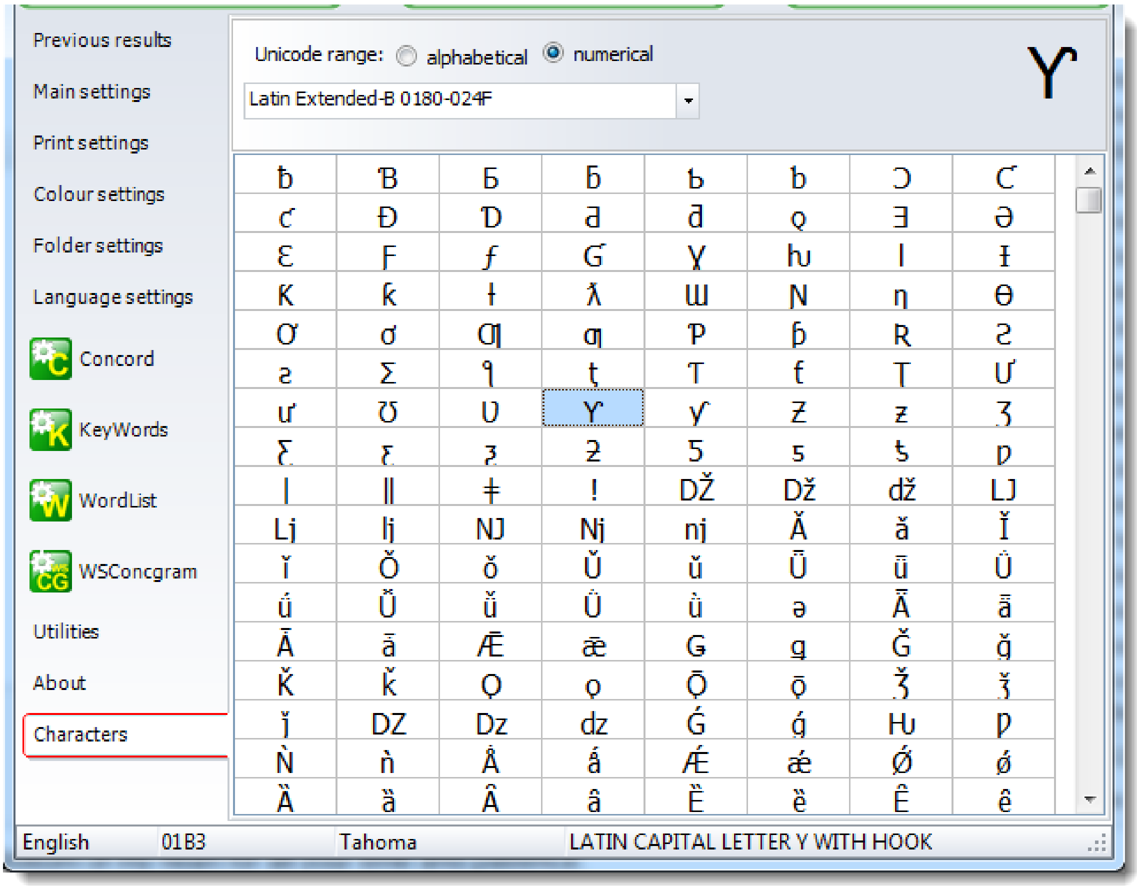
Tip

To check results after changing the code-page, select Choose Texts 44 and View the file in question. If you can't get it to look right, you've probably not got a cleaned-up plain text file but one straight from a word-processor. In that case, take it back into the word-processor (see here 444 for how to do that in MS Word) and save it as text again as a plain text file in Unicode.

See also: Text Formats 42, Choosing Accents & Symbols 420, Accented characters 419; Choosing Language 81

11.6.2 accents & symbols

When entering your search-word 163 you may need to insert symbols and accented characters into your search-word, exclusion word or context word, etc. If you have the right keyboard set for your version of Windows this may be very easy — if not, just choose the symbol in the main Controller 4 by clicking.



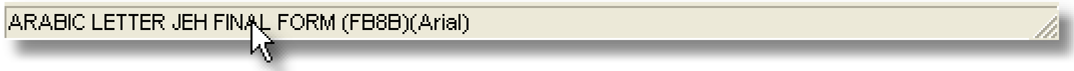
Unicode range: ○ alphabetical ● numerical

Latin Extended-B 0180-024F

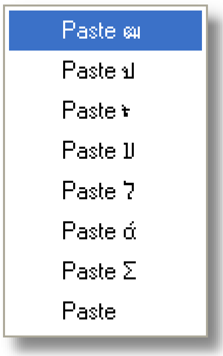
ƀ	Ɓ	Ƃ	ƃ	Ƅ	ƅ	Ɔ	Ƈ
ƈ	Ɖ	Ɗ	Ƌ	ƌ	ƍ	Ǝ	Ə
Ɛ	Ƒ	ƒ	Ɠ	Ɣ	ƕ	Ɩ	Ɨ
Ƙ	ƙ	ƚ	ƛ	Ɯ	Ɲ	ƞ	Ɵ
Ơ	ơ	Ƣ	ƣ	Ƥ	ƥ	Ʀ	Ƨ
ƨ	Ʃ	ƪ	ƫ	Ƭ	ƭ	Ʈ	Ư
ư	Ʊ	Ʋ	Ƴ	ƴ	Ƶ	ƶ	Ʒ
Ƹ	ƹ	ƺ	ƻ	Ƽ	ƽ	ƾ	ƿ
ǀ	ǁ	ǂ	ǃ	Ǆ	ǅ	ǆ	Ǉ
ǈ	ǉ	Ǌ	ǋ	ǌ	Ǎ	ǎ	Ǐ
ǐ	Ǒ	ǒ	Ǔ	ǔ	Ǖ	ǖ	Ǘ
ǘ	Ǚ	ǚ	Ǜ	ǜ	ǝ	Ǟ	ǟ
Ǡ	ǡ	Ǣ	ǣ	Ǥ	ǥ	Ǧ	ǧ
Ǩ	ǩ	Ǫ	ǫ	Ǭ	ǭ	Ǯ	ǯ
ǰ	Ǳ	ǲ	ǳ	Ǵ	ǵ	Ƕ	Ƿ
Ǹ	ǹ	Ǻ	ǻ	Ǽ	ǽ	Ǿ	ǿ
Ȁ	ȁ	Ȃ	ȃ	Ȅ	ȅ	Ȇ	ȇ

English 01B3 Tahoma LATIN CAPITAL LETTER Y WITH HOOK

Below, you will see which character has been selected


ARABIC LETTER JEH FINAL FORM (FB8B)(Arial)

with the current font (which affects which characters can be seen). You can choose a number of characters and then paste them into Concord, by right-clicking and choosing from the popup-menu:



These options above show Greek, Hebrew, Thai and Bengali characters have been clicked. The last one ("Paste") is the regular Windows paste.

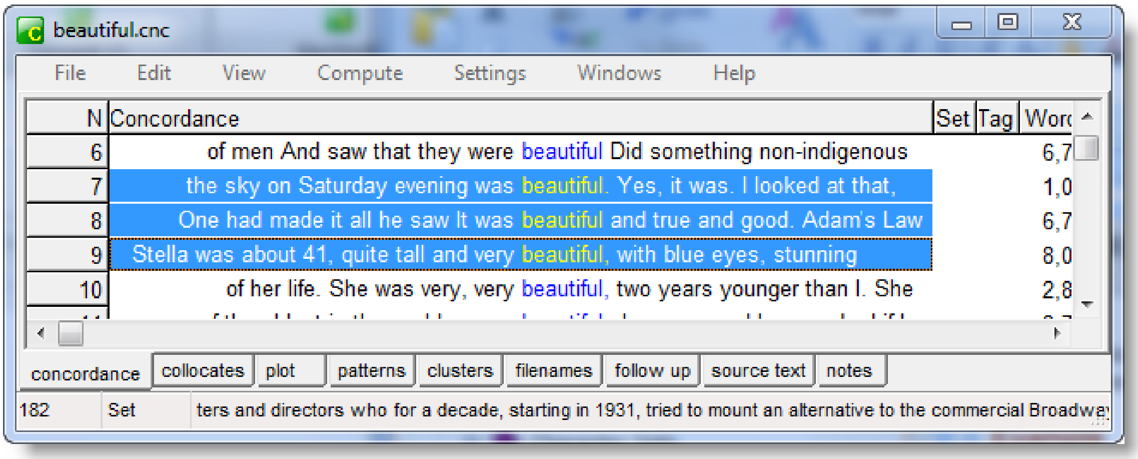
See also: Choosing Language 81, Change Language 419

11.7 clipboard

You can block an area of data, by using the cursor arrows and Shift, or the mouse, then press Ctrl +Ins or Ctrl+C to copy it to the clipboard. If you then go to a word processor, you can paste or ("paste special") the blocked area into your text. This is usually easier than saving as a text file 102 (or printing 97 to a file) and can also handle any graphic marks.

Example

1. Select some data. Here I have selected 3 lines of a concordance, just the visible text, no Set or Filenames information.

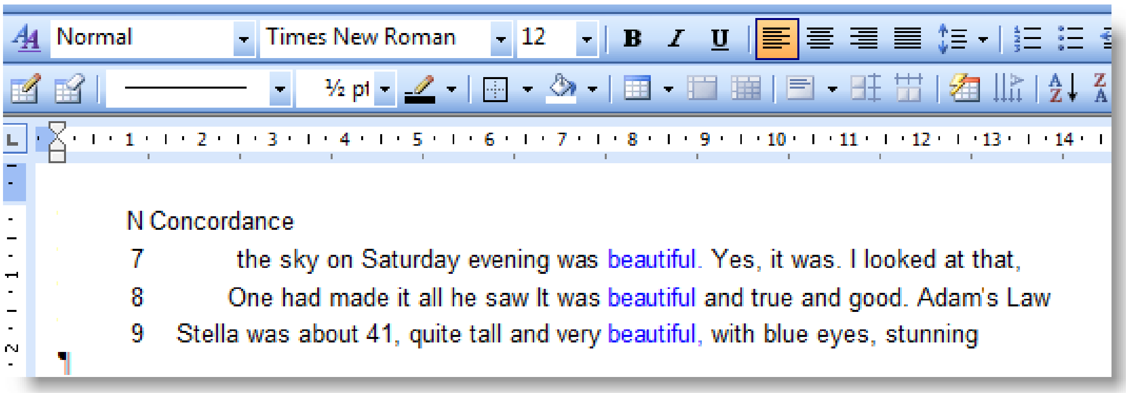


N	Concordance	Set	Tag	Word
6	of men And saw that they were beautiful Did something non-indigenous			6,7
7	the sky on Saturday evening was beautiful. Yes, it was. I looked at that,			1,0
8	One had made it all he saw It was beautiful and true and good. Adam's Law			6,7
9	Stella was about 41, quite tall and very beautiful, with blue eyes, stunning			8,0
10	of her life. She was very, very beautiful, two years younger than I. She			2,8

182 Set ters and directors who for a decade, starting in 1931, tried to mount an alternative to the commercial Broadwa

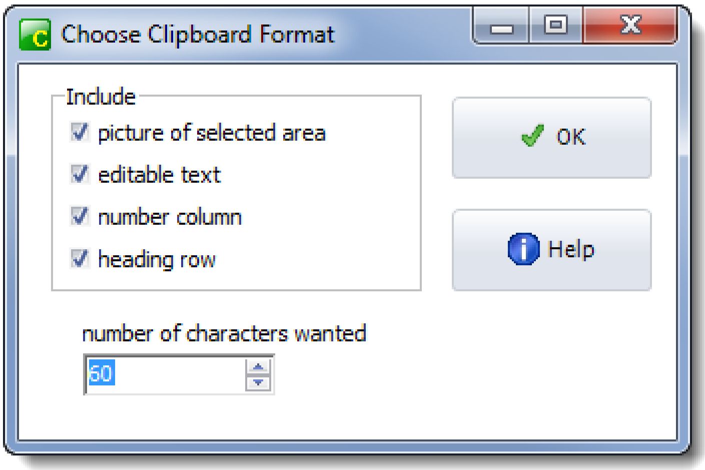
2. Hold down Control and press Ins or C.

In the case of a concordance, since concordance lines are quite complex, you will be asked whether you want a *picture* of the selected screen lines, which looks like this in MS Word:



N Concordance
7 the sky on Saturday evening was beautiful. Yes, it was. I looked at that,
8 One had made it all he saw It was beautiful and true and good. Adam's Law
9 Stella was about 41, quite tall and very beautiful, with blue eyes, stunning

with the colours and font resembling those in WordSmith, and/or plain text, and if so how many characters:



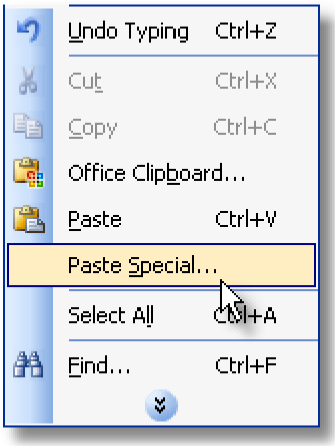
Include
☑ picture of selected area
☑ editable text
☑ number column
☑ heading row

number of characters wanted
60

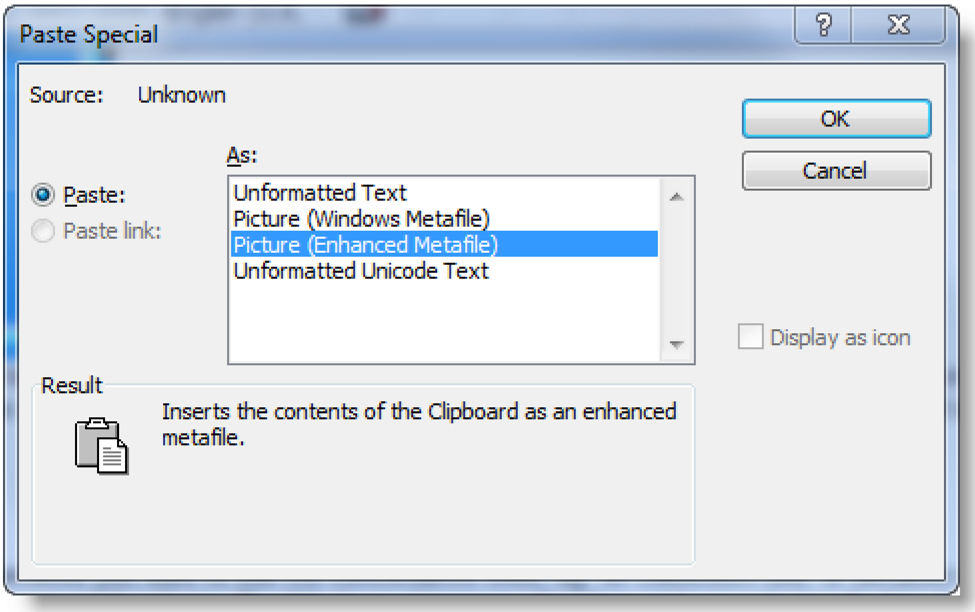
Once you've pressed OK, the data goes to the Windows "clipboard" ready for pasting into any other application, such as Excel, Word, Notepad, etc.

For all other types of lists, such as word-lists, the data are automatically placed in the Clipboard in both formats, as a picture and as text. You can choose either one and they will look quite different from each other!

Choose "Paste Special" in Word or any other application to choose between these formats.



and then, for the picture format



Paste Special

Source: Unknown

As:

● Paste:
○ Paste link:

- Unformatted Text
- Picture (Windows Metafile)
- Picture (Enhanced Metafile)
- Unformatted Unicode Text

□ Display as icon

Result

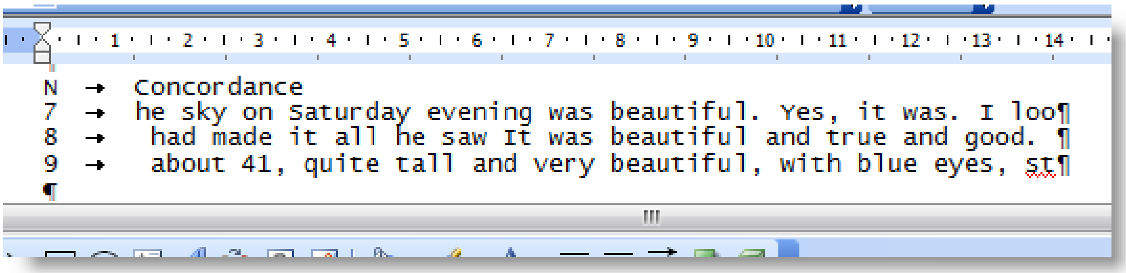
Inserts the contents of the Clipboard as an enhanced metafile.

You will probably use this picture format for your dissertation and will have to in the case of plotted data. In this concordance, you get only the words visible in your concordance line (not the whole line).
What you're pasting is a graphic which includes screen colours and graphic data. If you subsequently click on the graphic you will be able to alter the overall size of the graphic and edit each component word or graphic line (but not at all easily!). Note that if you select more lines than will subsequently fit on your page, MS Word may either shrink the image or only paste one pageful.

as plain text

Alternatively, you might want to paste as plain Unformatted Unicode text because you want to edit the concordance lines, eg. for classroom use, or because you want to put it into a spreadsheet such as MS Excel 102™. Here the concordance or other data are copied as plain text, with a tab between each column. The Windows plain text editor Notepad can only handle this data format. Microsoft Word will paste (using Shift+Ins or Ctrl+V) the data as text. It pastes in as many characters as you have chosen above, the default being 60.

At first, the concordance lines are copied, but they don't line up very nicely. Use a non proportional font, such as Courier or Lucinda Console, and keep the number of characters per line down to some number like 60 or so -- then it'll look like this:



```
N  →  Concordance
7  →  he sky on Saturday evening was beautiful. Yes, it was. I loo¶
8  →   had made it all he saw It was beautiful and true and good. ¶
9  →   about 41, quite tall and very beautiful, with blue eyes, st¶
¶
```

At 10 point text in Lucida Console, the width of the text with 60 characters and the numbers at the left comes to about 14 cm., as you can see To avoid word-wrapping, set the page format in Word to landscape, or keep the number of characters per line down to say 50 or 60 and the font size to

10.

avoid the heading and numbers in WordList or KeyWords too?

See advanced clipboard settings 36.

11.8 contact addresses

Downloads

You can get a more recent version at our website. There are also some free extra downloads (programs, word lists, etc.) there too. And links to sources of free text corpora.

Screenshots

visit http://lexically.net/wordsmith/support/get_started_guides.html for screenshots of what WordSmith Tools can do. This may give you useful ideas for your own research and will give you a better idea of the limitations of WordSmith too!

Purchase

Visit http://lexically.net/wordsmith/purchasing.htm for details of suppliers.

Complaints & Suggestions

Best of all, join Google Groups WordSmith Tools group and post your idea there so others can see the discussion. Or email me (mike (at) lexically.net). Please give me as full a description of the problem you need to tackle as you can, and details of the equipment too. Please don't include any attachments over 200K in size. I do try to help but cannot promise to…

11.9 date format

Date Format

Japanese date format year_month_day_hour_minute. At least it is logical, going from larger to smaller. Why aren't URLs organised in a logical order too?

11.10 Definitions

11.10.1 definitions

valid characters

Valid characters include all the characters in the language you are working with which are defined (by Microsoft) as "letters", plus any user-defined acceptable characters to be included within a word (such as the apostrophe or hyphen 435). That is, in English, **A, a,... Z, z** will be valid characters but **;** or @ or _ won't. In Greek, **δ** will count as a valid character. In Thai, (to patak) will be a valid character.

words

The word is defined as a *sequence of valid characters with a word separator 427 at each end.*

A word can be of any length, but for one to be stored in a word list, you may set the length you prefer (maximum of 50 characters) -- any which exceed your limit will get + tagged onto them at that point. You can decide whether or not to include words including numbers (e.g. **$35.50**) in text characteristics 124.

token and type

The term *token* is used to refer to running words and *type* to different words. So in **`This is my book, it is interesting`** we have 7 tokens but only 6 different types because **`is`** gets repeated.

clusters

A cluster is *a group of words which follow each other in a text*. The term phrase is not used here because it has technical senses in linguistics which would imply a grammatical relation between the words in it. In WordList cluster processing 278 or Concord cluster processing 175 there can be no certainty of this, though clusters often do match phrases or idioms. See also: general cluster information 448.

sentences

The sentence is defined as *the full-stop, question-mark or exclamation-mark (.?!) and (equivalents in languages such as Arabic, Chinese, etc.) immediately followed by one or more word separators 427 and then a number or a currency symbol, or a letter in the current language which isn't lower-case*. Note: languages which do not distinguish between lower-case and upper-case characters do not technically count any as lower case or upper case. (For more discussion see Starts and Ends of Text Segments 146 or Viewer & Aligner technical information 390.)

paragraphs

Paragraphs are user-defined. See Starts and Ends of Text Segments 146 for further details.

headings

Headings are also user-defined -- see Starts and Ends of Text Segments 146.

texts

A text in WordSmith means what most non-linguists would call a text. In a newspaper, for example, there might be 6 or 7 "texts" on

each page. This also means that a text = a file on disk. If it doesn't you're better off totally ignoring the "Texts" column in any WordSmith output.

chargrams

A chargram is a sequence of N consecutive valid characters (excluding digits and punctuation) found in text. e.g. **`ABI,ABL,ABO`** etc. In English the most frequent 3-chargrams are **`THE, ING, AND, ION`**.

See also: Setting Text Characteristics 124, Keyness 236, Key key-word 240, Associate 243

11.10.2 word separators

Conventionally one assumes that one word is distinguished from the next by the presence of spaces at either end. But WordSmith Tools also includes within word separators certain standard codes used by most word processors: page eject code (12), tabs (9), carriage return (13) and line feed (10), end-of-text (26). Besides, hyphens 435 may optionally be considered to split words like self-access into two words.

Note that in Chinese and Japanese which do not separate words in this way, any WordSmith functions which require word-separation will not work unless you get your texts previously tagged with word-separators.

11.11 demonstration version

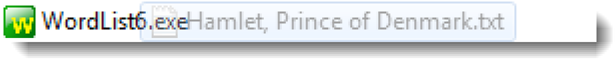
The demonstration version of **WordSmith Tools** offers *all* the facilities of the complete suite, except that any screen which shows a list (of words in a word-list, or concordance lines, etc.) is limited to a small number of lines which can be shown or printed. (If you save data, all of it will be saved; it's just that you can't see it all in the demo version.)

See also: Installing 21, Version Information 451, Contact Addresses 425.

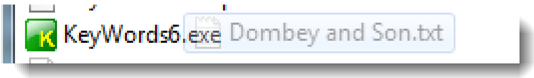
11.12 drag and drop

You can get WordSmith to compute some results simply by dragging.

If you have **WordList** open you can simply drag a text file onto it from Windows Explorer and it will create a word-list there and then using default settings. Or if it is not open, drag your text file to the `WordList6.exe` file. Here, *Hamlet* is being dragged onto the WordList tool.


WordList6.exe Hamlet, Prince of Denmark.txt

If you have **KeyWords** open you can simply drag a text file onto it from Windows Explorer. If you have a valid word list set as the reference corpus, it will compute the key words.
Or if it is not open, drag your text file to the `KeyWords6.exe` file, as in this screenshot where the Dickens novel *Dombey and Son.txt* is being dragged onto the KeyWords file.


KeyWords6.exe Dombey and Son.txt

If you drag a word-list made by WordList (`.LST` ending), a concordance (`.CNC`), a key word list (`.KWS`) etc. onto the Controller 4, it will open it with the appropriate tool.

11.13 edit v. type-in mode

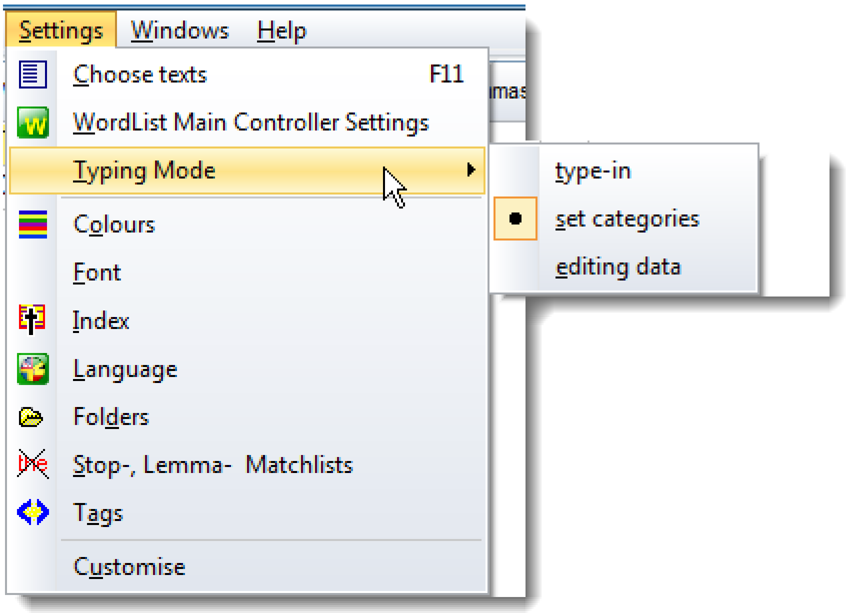
Most windows allow you to press keys either

- to edit your data (edit mode), or
- to get quickly to a place in a list (type-in mode).

Concordance windows use key presses also for setting categories 168 for the data, or for blanking 168 out the search word.

In type-in mode, your key-presses are supposed to help you get quickly 109 to the list item you're interested in, e.g by typing `theocr` to get to (or near to) `theocracy` in a word list. If you've typed in 5 letters and a match is found, the search stops.

Changing mode is done in the menu: Settings | Typing Mode:



See also: user-defined categories 168.

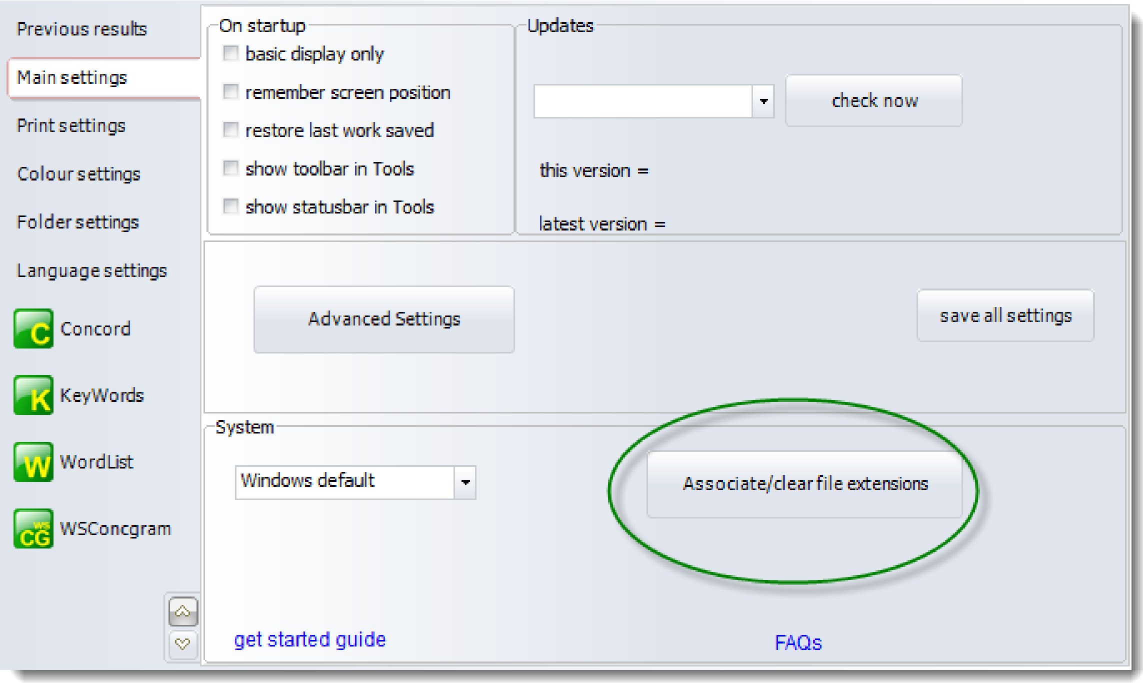
11.14 file extensions

The standard file-extensions used in WordSmith are

`.cnc`	concordance file
`.lst`	word list
`.mut`	mutual information list
`.dcl`	detailed consistency list

`.tokens, .types`	word list index file
`.kws`	key words file
`.kdb`	key word database file
`.base_pairs, .base_index_cg`	WSConcgram files
`.ali`	aligner list
`.vwr`	viewer list

In the Controller's Main settings, or on installing, you can if you wish associate (or disassociate) the current file-types with WordSmith in the Registry. The advantage of association is that Windows will know what Tool to open your data files with.



On startup
- □ basic display only
- □ remember screen position
- □ restore last work saved
- □ show toolbar in Tools
- □ show statusbar in Tools

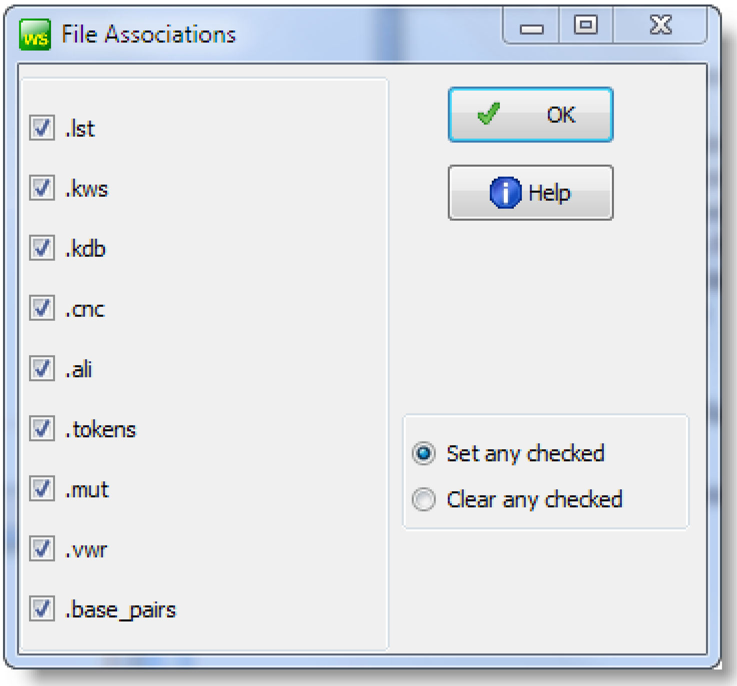
Updates

this version =

latest version =

System

Windows default



File Associations

☑ .lst

☑ .kws

☑ .kdb

☑ .cnc

☑ .ali

☑ .tokens

☑ .mut

☑ .vwr

☑ .base_pairs

● Set any checked

○ Clear any checked

11.15 finding source texts

For some calculations the original source texts need to be available. For example, for Concord to show you more context 165 than has been saved for each line, it'll need to re-read the source text. For KeyWords to calculate a dispersion plot 251, it needs to look at the source text to find out which KWs came near each other and compute positions of each KW in the text and KW links 247.

If you have moved or deleted the source file(s) in the meantime, this won't be possible.

See also : Source texts 116, Editing filenames 110, Choosing source files 44, find files 269.

11.16 flavours of Unicode

What is Unicode?

What WordSmith requires for many languages (Russian, Japanese, Greek, Vietnamese, Arabic etc.) is Unicode. (Technically UTF16 Unicode, little-endian.) It uses 2 bytes for each character. One byte is not enough space to record complex characters, though it will work OK for the English alphabet and some simple punctuation and number characters.

UTF8, a format which was devised for many languages some years ago when disk space was limited and character encoding was problematic, is in widespread use but is generally *not suitable*. That's because it uses a variable number of bytes to represent the different characters. A to Z will be only 1 byte but for example Japanese characters may well need 2, 3 or even more bytes to represent one character.

There are a number of different "flavours" of Unicode as defined by the Unicode Consortium.

MS Word offers

- Unicode
- Unicode (Big-Endian) (generated by some Mac or Unix software)
- Unicode (UTF-7)
- Unicode (UTF-8)

The last two are 1-byte versions, not really Unicode in my opinion. WordSmith wants the first of these but should automatically convert from any of the others. If you are converting text 365, prefer Unicode (little-endian), UTF16.

Technical Note

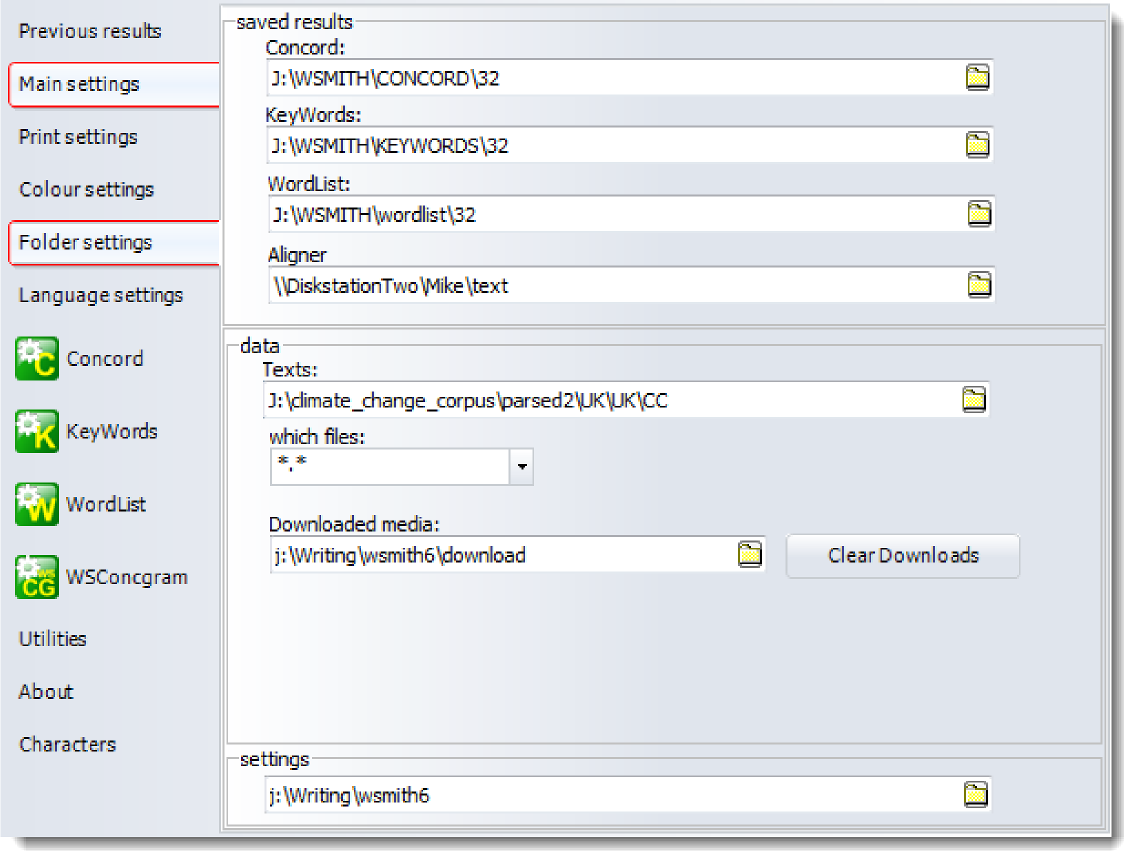
There are other flavours too and there is much more complexity to this topic than can be explained here, but essentially what we are trying to achieve is a system where a character can be stored in the PC in a fixed amount of space and displayed correctly.

Precomposed

In a few cases in certain languages, some of your texts may have been prepared with a character followed by an accent, such as **A** followed by **^** where the intention is for the software to display them merged (**Â**), instead of using precomposed characters where the two are merged in the text file. See the explanation in Advanced Settings 36 if you need to handle that situation.

11.17 folders\directories

Found in main Settings menu in all Tools. Default folders can be altered in WordSmith Tools or set as defaults 113 in **wordsmith6.ini**.



saved results

Concord: J:\WSMITH\CONCORD\32

KeyWords: J:\WSMITH\KEYWORDS\32

WordList: J:\WSMITH\wordlist\32

Aligner: \\DiskstationTwo\Mike\text

data

Texts: J:\climate_change_corpus\parsed2\UK\UK\CC

which files: *.*

Downloaded media: j:\Writing\wsmith6\download

settings

j:\Writing\wsmith6

- Concordance Folder: for your concordance files.
- KeyWords Folder: for your key-word list files.
- WordList Folder: where you will usually save 101 your word-list files.
- Aligner: for your dual-text aligned work 381
- Texts Folder: where your text files are to be found.
- Downloaded Media: where your sound & video files 212 will be stored after downloading the first time from the Internet.
- Settings: where your settings files (.ini files and some others) are kept.

If you write the name of a folder which doesn't exist, WordSmith Tools will create it for you if possible. (On a network, this will depend on whether you have rights to create folders and save 101 files.)

If you change your Settings folder, you should let WordSmith copy any **`.ini`** and other settings files which have been created so that it can keep track of your language preferences, etc.

Note: in a network, drive letters such as **`G:, H:, K:`** change according to which machine you're running from, so that what is **G:\texts\my text.txt** on one terminal may be **H:\texts\my text.txt** on another. Fortunately network drives also have names structured like this: \\computer_name\drive_name\. You will find that these names can be used by **WordSmith**, with the advantage that the same text files can be accessed again later.
If you run WordSmith from an external hard drive or a flash drive 21, where again the drive letter may change, you will find WordSmith arranges that if your folders are on that same drive they will change drive letter automatically once you have saved your defaults 113.

Tip

Use different folders for the different functions in WordSmith Tools. In particular, you may end up making a lot of word lists and key word lists if you're interested in making databases 237 of key words. It is theoretically possible to put any number of files into a folder, but accessing them seems to slow down after there are more than about 500 in a folder. Use the batch facility to produce very large numbers of word list or key words files. I would recommend using a **`\keywords`** folder to store **`.kdb`** files, and **`\keywords\genre1, \keywords\genre2,`** etc. for the **`.kws`** files for each genre.

See also: finding source texts 430.

11.18 formulae

For computing collocation strength, we can use

- the joint frequency of two words: how often they co-occur, which assumes we have an idea of how far away counts as "neighbours". (If you live in London, does a person in Liverpool count as a neighbour? From the perspective of Tokyo, maybe they do. If not, is a person in Oxford? Heathrow?)
- the frequency word 1 altogether in the corpus
- the frequency of word 2 altogether in the corpus
- the span or horizons 180 we consider for being neighbours
- the total number of running words in our corpus: total tokens

Mutual Information

Log to base 2 of (A divided by (B times C))

where

A = joint frequency divided by total tokens
B = frequency of word 1 divided by total tokens
C = frequency of word 2 divided by total tokens

MI3

Log to base 2 of ((J cubed) times E divided by B)

where

J = joint frequency
F1 = frequency of word 1
F2 = frequency of word 2
E = J + (total tokens-F1) + (total tokens-F2) + (total tokens-F1-F2)
B = (J + (total tokens-F1)) times (J + (total tokens-F2))

T Score

(J - ((F1 times F2) divided by total tokens)) divided by (square root of (J)) where

J = joint frequency
F1 = frequency of word 1
F2 = frequency of word 2

Z Score

(J - E) divided by the square root of (E times (1-P))

where

J = joint frequency
S = collocational span
F1 = frequency of word 1
F2 = frequency of word 2
P = F2 divided by (total tokens - F1)
E = P times F1 times S

Dice Coefficient

(J times 2) divided by (F1 + F2)

where

J = joint frequency
F1 = frequency of word 1 or corpus 1 word count
F2 = frequency of word 2 or corpus 2 word count
Ranges between 0 and 1.

Log Likelihood

based on Oakes 417 p. 170-2.

2 times (

a Ln a + b Ln b + c Ln c + d Ln d

- (a+b) Ln (a+b)

- (a+c) Ln (a+c)

- (b+d) Ln (b+d)

- (c+d) Ln (c+d)

+ (a+b+c+d) Ln (a+b+c+d)

)

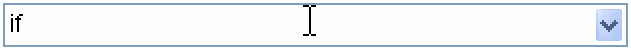
where

a = joint frequency
b = frequency of word 1 - a
c = frequency of word 2 - a
d := frequency of pairs involving neither word 1 nor word 2
and "Ln" means Natural Logarithm

See also: this link from Lancaster University, Mutual Information 289

11.19 history list

History List: many of the combo-boxes in WordSmith like this one for choosing a search-word remember what you type in so you can look them up by pressing the down arrow at the right.



if

11.20 HTML, SGML and XML

These are formats for text exchange. The most well known is HTML, Hypertext Markup Language, used for distributing texts via the Internet. SGML is Standard Generalized Markup Language, used by publishers and the BNC; XML is Extensible Markup Language, intermediate between the other two.

All these standards use plain text with additional extra tags, mostly angle-bracketed, such as <h1> and </h1>. The point of inserting these tags is to add extra sorts of information to the text:

1 a header (**<head>**) supplying details of the authorship & edition

2 how it should display (e.g. **<bold>, <italics>**)

3 what the important sections are (**<h1>** marks a heading, **<body>** is the body of the text)

4 how special symbols should display (&eacute corresponds to é)

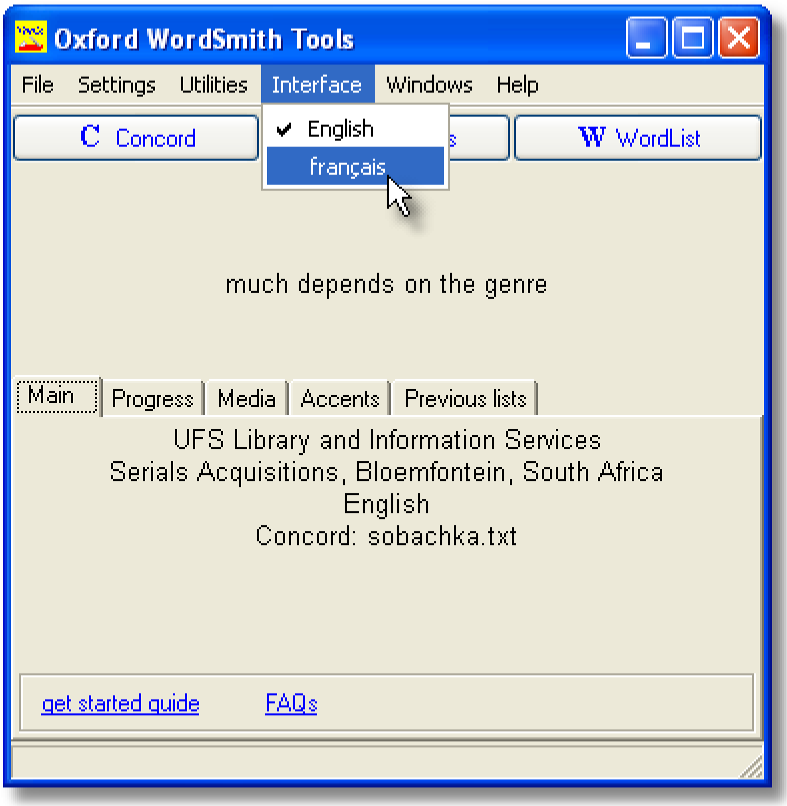
See also: Overview of Tags 131

11.21 hyphens

The character used to separate words. The item "self-help" can be considered as 2 words or 1 word, depending on Language Settings 124.

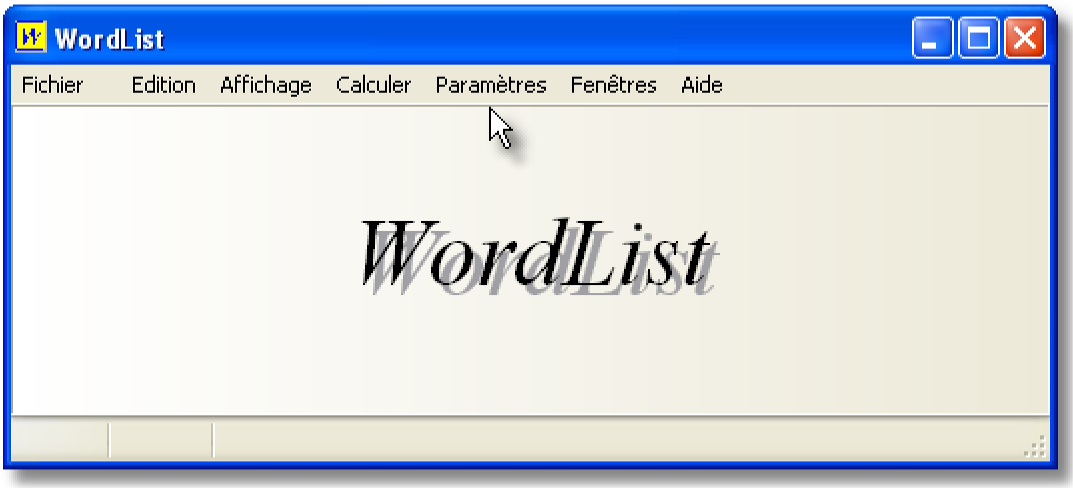
11.22 international versions

WordSmith can operate with a series of interfaces depending on the language chosen.


Oxford WordSmith Tools
File Settings Utilities Interface Windows Help
✓ English
français
C Concord
W WordList
much depends on the genre
Main Progress Media Accents Previous lists
get started guide
FAQs

UFS Library and Information Services
Serials Acquisitions, Bloemfontein, South Africa
English
Concord: sobachka.txt

If you choose French this is what you see in all of WordSmith.



See also: acknowledgements 416

11.23 limitations

The programs in WordSmith Tools can handle virtually unlimited amounts of text. They can read text from CD-ROMs, so giving access to corpora containing many millions of words. In practice, the limits are reached by a) storage 447 and b) patience.

You can have as many copies of each Tool running at any one time as you like. Each one allows you to work on one set of data.

Tags to ignore 132 or ones containing an asterisk can span up to 1,000 characters.

When searching for tags to determine whether your text files meet certain requirements 137, only the first 2 megabytes of text are examined. For Ascii that's 2 million characters, for Unicode 1 million.

Tip

Press F9 to see the "About" box -- it shows the version date and how much memory 447 you have available. If you have too little memory left, try a) closing down some applications, b) closing WordSmithTools and re-entering.

See also: Specific Limitations of each Tool 437

11.23.1 tool-specific limitations

Concord limitations

You can compute a virtually unlimited number of lines of concordance using Concord.

Concord allows 80 characters for your search-word or phrase 159, though you can specify an unlimited number of concordance search-words in a search-word file 161.

Each concordance can store an unlimited number of collocates with a maximum horizon 180 of 25 words to left and right of your search-word.

WordList limitations

A head entry can hold thousands of lemmas 270, but you can only join up to 20 items in one go using F4. Repeat as needed.

Detailed Consistency 263 lists can handle up to 50 files.

KeyWords limitations

One key-word plot per key-word display. (If you want more, call up the same file in a new display window.)

number of link 247-windows per key-word plot 251 display: 20.

number of windows of associates 241 per key key-word display: 20.

File Utilities: Splitter limitations

Each line of a large text file can be up to 10,000 characters in length. That is, there must be an <Enter> from time to time!

Text Converter limitations

There can be up to 500 strings to search-and-replace for each.

Each search-string and each replace-string can be up to 80 characters long.

An asterisk must not be the first or last character of the search-string.

When the asterisk is used to retain information, the limit is 1,000 characters.

Viewer & Aligner limitations

If you choose the View option ▤ when choosing texts, Viewer & Aligner will call up the first 10 source text files selected.

When choosing texts or jumping into the middle of a text (e.g. after choosing ▤ in Concord), Viewer & Aligner will only process 10,000 characters of each file, to speed things up in the case of very large files, but you can get it to "re-read" the file by pressing □ to refresh the display, after which it will read the whole text.

See also: General Limitations 437

11.24 links between tools

Linkage with Word Processors, Spreadsheets etc.

All the windows showing lists or texts can easily copy selected information to the clipboard 422. (Use Ctrl+Ins or Ctrl/C to insert).

W Where you see this symbol, you can send any selected data straight to a new Microsoft Word™ document.
Where you see an URL (such as http://lexically.net) you can click to access your browser.

Links between the various Tools

The programs in **WordSmith Tools** are linked to each other via wordsmith.exe 4 (the one which says " WordSmith Tools Controller 4" in its caption, and is found in the top-left corner of your screen). This handles all the defaults 113, such as colours, folders, fonts, stop lists, etc.

In general, if you press Ctrl+C in **WordList** or **KeyWords** you'll go straight to a concordance, computed using the current word and using the current files.

Each Tool will send as much relevant information as possible to the Tool being called. This will include: the current word (the one highlighted in the scrolling window) and the text files where any current information came from.

Example: after computing a word list based on 3 business texts, you discover that the word *hopeful* is more frequent than you had expected. You want to do a concordance on that word, using the same texts. Place the highlight on *hopeful,* hold down Control and press C. Now you can see whether *hopeful* is part of a 3-word cluster 175, or view a dispersion plot.
Example: after computing a key words database 237 using 300 business texts, you discover that

the word *bid* seems to be a key key-word, and that it's associated with *company*, *shares* etc. Place the highlight on *bid,* press Control-C and a concordance will be computed using the same 300 texts. Now you can check out the contexts: is *bid* a bid for power, or is it part of a tendering process?
Example: you have a concordance of *green*. Now press Control-W to generate a word list of the same text files. Press Control-K to compare this word list with a reference corpus list to see what the key words are in these text files.

11.25 keyboard shortcuts

scrolling windows:

Control+Home	to top of scrollable list
Control+End	to end of scrollable list
if it's ordered alphabetically:	type-in your search-word[109]
and if it scrolls horizontally:	Home -- to left edge End -- to right edge

hotkeys:

Shift-cursor keys	block a section

F1	help
Ctrl+F2	save results[101]
F3	print preview
Ctrl+P	print results
F4	join entries
Ctrl+F4	unjoin
Alt+F5	mark entries for joining
Shift+Alt+F5	unmark entry
F5	refresh a list
Shift+Ctrl+F8	auto set row height in Concord
F6	re-sort[315]
Ctrl+F6	reverse word sort[315]
Shift+Ctrl+F6	word-length sort
F7	view source text
F8	grow line height
Ctrl+F8	shrink
F9	About box (shows version-date and memory availability)
F10	compute collocates
F11	choose texts
Ctrl+Shift+C	compute concordance
Ctrl+C	copy

Ctrl+F3	find again
Alt+D	find next deleted entry
Ctrl+L	layout & columns of data
Ctrl+M	play media file
Ctrl+N	new
Ctrl+U	undo
Ctrl+V	paste
Ctrl+W	close
Alt+X	e**X**it the Tool
Ctrl+Z	Zap[129] deleted lines
Del	delete
Numeric -	delete to the end
Ins	restore deleted entry
Numeric +	restore to the end

see also: Menu items and Buttons[441]

11.26 machine requirements

This version of **WordSmith Tools** is designed for machines with:

- at least 1GB of RAM
- at least 200MB of hard disk space
- Windows™ XP or later, or an emulator of one of these if using an Apple Mac or Unix system.

You will find it runs better on a faster[448] machine, especially if there's plenty of RAM[447].
You can run WordSmith from a memory stick[21] on a fast computer better than on a slow computer. (You can run WordSmith on a tiny 10" screen laptop with Windows Starter and little power but all applications on those are slow and there is not much screen for your results.)

There is no Apple Mac version but see http://lexically.net/wordsmith/mac_intel.htm for details on how to use WordSmith on a Mac.

11.27 manual for WordSmith Tools

This help file exists in the form of a manual, which you get when you install[21]. The file (**wordsmith.pdf**), is in Adobe Acrobat™ format. It has a table of contents and a fairly detailed index (which I used **WordList** and **KeyWords** to help me create). Most people find paper easier to deal with than help files!

You may find it useful to see screenshots of **WordSmith** in action: ideas are listed here[425].

11.28 menu and button options

These functions may or may not be visible in each Tool depending on the capacity of the Tool or the current window of data -- the one whose caption bar is highlighted.

advanced

allows access to advanced features

associates

opens a new window showing Associates 241.

auto-join

joins (lemmatises 270) automatically.

auto-size

re-sizes each line of a display so that each one shows as much data as it should. Most windows have lines of a fixed size but some, e.g. in Viewer, allow you to adjust row heights. This adjusts line heights according to the current highlighted column of data.

close (Ctrl+W)

closes a window of data

clumps

computes clumps 243 in a keywords database

regroup clumps

regroups 244 the clumps

clusters

computes concordance clusters 175.

collocates

shows collocates 179 using concordance data.

compute

calculates a new column of data 63 based on calculator functions and/or existing data.

redo collocates

recalculates collocates, e.g. after you've deleted concordance lines.

column totals

computes totals, min, max, mean, standard deviation for each column 62 of numerical data.

concordance (Shift+Ctrl+C)

within KeyWords, WordList, starts Concord and concordances the highlighted word(s) using the original source text(s).

copy (Ctrl+C)

allows you to copy 66 your data to a variety of different places (the printer, a text file 102, the clipboard 422, etc.).

edit

allows editing 72 of a list or searches for a word (type-in search 109).

exit (Alt+X)

quits a Tool.

edit or type-in mode

alternates between edit and type-in mode.

filenames

opens a new window showing the file names from which the current data derived. If necessary you can edit them 113.

find files

finds any text files which contain all the words you've marked.

grow

increases the height of all rows to a fixed size. See shrink (▬) below.

help (F1)

opens WordSmith Help (this file) with context-sensitive help.

join

joins one entry to another e.g. sentences in Viewer, words in WordList (lemmatisation 270).

layout

This allows you to alter many settings for the layout 87: the colour of each column, whether to hide a column of data, typefaces and column widths.

links

computes links 247 between words in a key-words plot.

mark

marks an entry for joining 270 or finding files 75.

match lemmas

checks each item in the list against ones from a text file of lemmatised forms and joins 270 any that match.

match list

matches up the entries in the current list against ones in a "match list file" or template 92, marking any found with (~).

relation

computes mutual information 289 or similar scores in a WordList index list 276.

new... (Ctrl+N)

gets you started 2 in the various Tools, e.g. to make a concordance, a word list, or a key words list.

open... (Ctrl+O)

gives you a chance to choose a set of saved results.

patterns

computes collocation patterns 207.

play media (Ctrl+M)

plays a media file 212.

plot

opens a new window showing a Concord dispersion plot 191 or KeyWords plot 251.

print preview (F3)

previews your window data for printing (Ctrl+P); can print to file, which is equivalent to "save as text 102".

redo

undoes an undo.

refresh (F5)

re-draws the screen (in Viewer re-reads your text file).

remove duplicates

removes any duplicate concordance 207 lines.

replace

search & replace, e.g. to replace drive or folder data, when editing file-names 113 where the source texts have been moved.

re-sort

re-sorts lists (e.g. in frequency as opposed to alphabetical order) in Concord 208, KeyWords 252 or WordList 315.

ruler

shows/hides vertical divisions in any list; text divisions in a KeyWords plot 251. Click ruler in a menu to turn on or off or change the number of ruler divisions for a plot 191.

save (also Ctrl+F2)

saves your data 101 using existing file-name; if it's a new file asks for file-name first.

save as

saves after asking you for a file-name.

save as text

saves as a .txt file: plain text.

search

searches 109 within a list.

shrink

reduces the height of all rows to a smaller fixed height. See grow (🟪) above.

statistics

shows detailed statistics 298.

statusbar

toggles on & off the "status bar" (at the bottom of a window, shows comments and the status of what has been done).

summary statistics

opens a new window showing summary statistics 66, e.g. proportion of lemmas to word-types.

toolbar

toggles on & off a toolbar with the same buttons on it as the ones you chose when you customised popup menus 31.

undo (Ctrl+U)

undoes last operation.

unjoin

unjoins any entries that have been joined, e.g. lemmatised 270 entries.

view source text

shows the source text 379 and highlights any words currently selected in the list.

, Microsoft Excel or Word™

save formatted data for Excel or Word.

wordlist

within KeyWords, makes a word list 249 using the current data.

zap (Ctrl+Z)

zaps 129 any deleted entries.

see also: Keyboard Shortcuts 439, Customising popup menus 31.

11.29 MS Word documents

Inside a **.doc** or **.docx** file there is a lot of extra coding apart from the plain text words. (Actually, a **.docx** doesn't even seem to show the ordinary text words inside it!) For example, the name of your printer, the owner of the software, information about styles etc. For accurate results, WordSmith needs to use clean text where these have been removed.

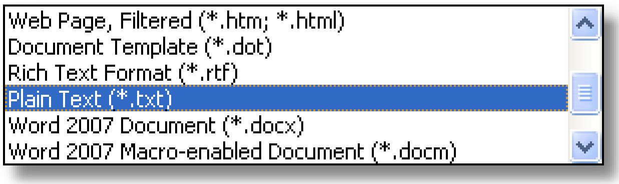
converting your .DOC or .DOCX files

The easiest method, for multiple **.doc** or **.docx** files, is to convert using the Text Converter 365.

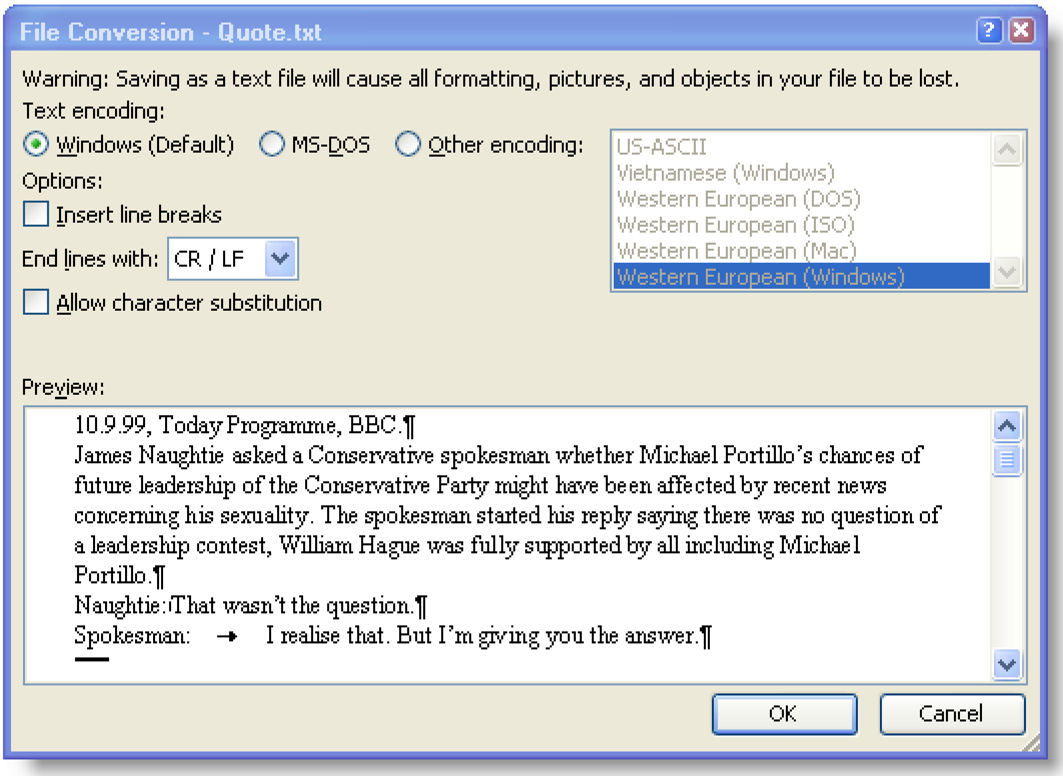
Alternatively you can do it in Word

To convert a **.doc** or **.docx** into plain text in Word can be done thus:

Chose *File* | *Save As* | *Plain text:*


Web Page, Filtered (*.htm; *.html)
Document Template (*.dot)
Rich Text Format (*.rtf)
Plain Text (*.txt)
Word 2007 Document (*.docx)
Word 2007 Macro-enabled Document (*.docm)

then choose Windows (1-byte per character)



File Conversion - Quote.txt

Warning: Saving as a text file will cause all formatting, pictures, and objects in your file to be lost.

Text encoding:

● Windows (Default) ○ MS-DOS ○ Other encoding:

US-ASCII
Vietnamese (Windows)
Western European (DOS)
Western European (ISO)
Western European (Mac)
Western European (Windows)

Options:

□ Insert line breaks

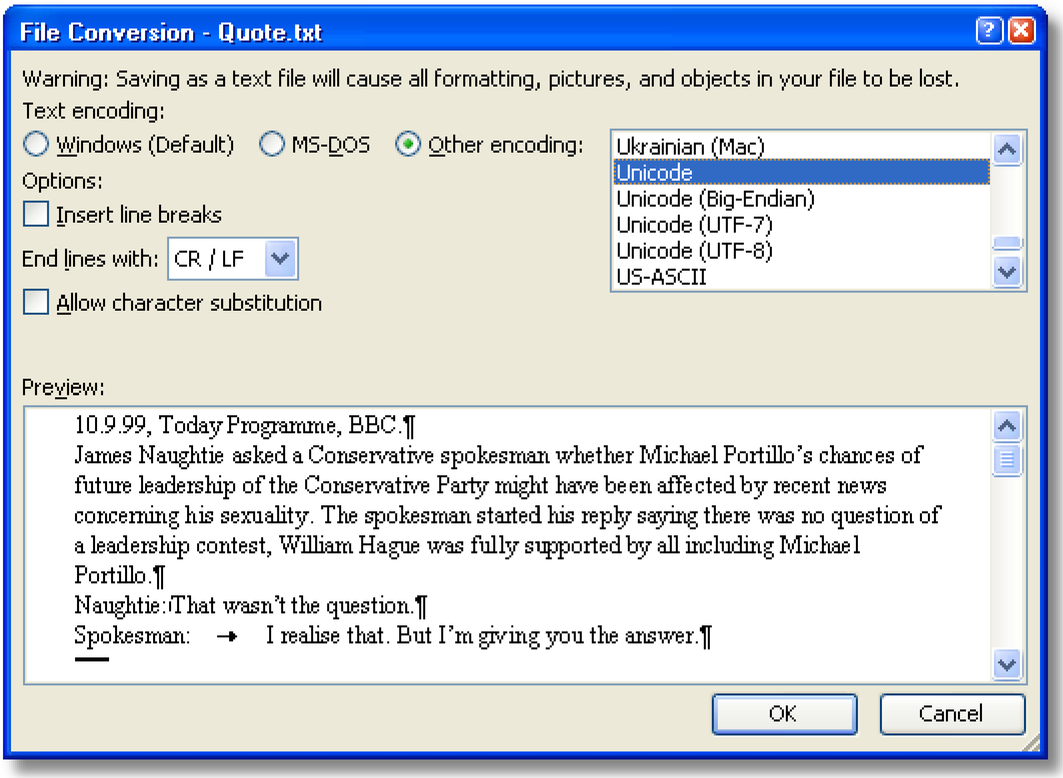
End lines with: CR / LF

□ Allow character substitution

Preview:

10.9.99, Today Programme, BBC.¶
James Naughtie asked a Conservative spokesman whether Michael Portillo's chances of future leadership of the Conservative Party might have been affected by recent news concerning his sexuality. The spokesman started his reply saying there was no question of a leadership contest, William Hague was fully supported by all including Michael Portillo.¶
Naughtie: That wasn't the question.¶
Spokesman: → I realise that. But I'm giving you the answer.¶

or Other encoding -- Unicode (2-bytes):



File Conversion - Quote.txt

Warning: Saving as a text file will cause all formatting, pictures, and objects in your file to be lost.

Text encoding:

○ Windows (Default) ○ MS-DOS ● Other encoding:

Ukrainian (Mac)
Unicode
Unicode (Big-Endian)
Unicode (UTF-7)
Unicode (UTF-8)
US-ASCII

Options:

□ Insert line breaks

End lines with: CR / LF

□ Allow character substitution

Preview:

10.9.99, Today Programme, BBC.¶
James Naughtie asked a Conservative spokesman whether Michael Portillo's chances of future leadership of the Conservative Party might have been affected by recent news concerning his sexuality. The spokesman started his reply saying there was no question of a leadership contest, William Hague was fully supported by all including Michael Portillo.¶
Naughtie: That wasn't the question.¶
Spokesman: → I realise that. But I'm giving you the answer.¶

11.30 never used WordSmith before

For users who are starting out with WordSmith for the first time, the whole process can seem complex. (After all, the first time you used word-processing software that seemed tricky -- but you already knew what a text is and how to write one...)

So a small text file accompanies the WordSmith installation, and if WordSmith thinks you have never used it before, it will automatically choose that text file for you to start using Concord, WordList etc. WordSmith's method of knowing that you are a new user is

1) have any concordances or wordlists been saved 101?

and

2) has no set of favourite text 50 files been saved for easy retrieval?

11.31 numbers

Depending on Language and Text Settings 124, you might wish to include or exclude numbers from word lists.

11.32 plot dispersion value

The point of it

A dispersion value is the degree to which a set of values are uniformly spread. Think of rainfall in the UK -- generally fairly uniformly spread throughout the year. Compare with countries which have a rainy season.
In linguistic terms, one might wish to know how the occurrences of a word like *skull* are distributed in Hamlet, and WordSmith has shown this in plot form since version 1. The dispersion value statistic gives mathematical support to this and makes comparisons easier.

How it is calculated

The plot dispersion calculated in KeyWords and Concord dispersion plots uses the first of the 3 formulae supplied in Oakes 417 (1998: 190-191), which he reports as having been evaluated as the most reliable.
Like the ruler 441, it divides the plot into 8 segments for this.
It ranges from 0 to 1, with 0.9 or 1 suggesting very uniform dispersion and 0 or 0.1suggesting "burstiness" (Katz 417, 1996)

See also: KeyWords plot 251, Concord dispersion plot 191.

11.33 RAM availability

The more RAM (chip memory) you have in your computer, the faster it will run and the more it can store. As it is working, each program needs to store results in memory. A word list of over 80,000 entries, representing over 4 million words of text, will take up roughly 3 Megabytes of memory. (In Finnish it would be much more.) When memory is low, Windows will attempt to find room by putting some results in temporary storage on your hard disk. If this happens, you'll probably hear a lot of clicking as it puts data onto the disk and then reads it off again. You will probably hear *some* clicking anyway as most of the programs in **WordSmith Tools** access your original texts from the hard disk, but a constant barrage of *thrashing* shows you've reached your machine's natural limits.

You can find out how much storage you have available even in the middle of a process, by pressing F9 (the About option in the main *Help* menu of each program). The first line states the RAM availability. The other figures supplied concern Windows system resources: they should not be a problem but if they do go below about 20% you should save results 101, exit Windows and re-enter.

Theoretically, word lists and key word lists can contain up to 2,147,483,647 separate entries. Each of these words can have appeared in your texts up to 2,147,483,647 times. (This strange number 2,147,483,647, half of 2 to the power 32, is the largest signed integer which can be stored in 32 bits and is also called 2 Gigabytes.) You are not likely to reach this theoretical limit: for the item *the* to have occurred 2,147,483,647 times in your texts, you would have processed about 30 thousand million words (1 CD-ROM, containing only plain text, can hold about 100 million words so this number represents some 300 CD-ROMs.) You would have run out of RAM long before this.

If you have a Gigabyte of RAM or more you should be able to have a copy of a word-list based on millions of words of text, and at the same time have a powerful word-processor and a text file in memory.

See also: speed 448

11.34 reference corpus

Reference Corpus

A corpus of text which you use for comparative purposes. For example, you might want to compare a given piece of text with the British National Corpus, a collection of 100 million words. Useful when computing key words 229.

In the Controller 4 you can set your reference corpus word list 113 for KeyWords and Concord to make use of. (That is, a word list 318 created using the WordList 258 tool.)

11.35 restore last file

By default, the last word list, concordance or key words listing that you saved or retrieved will be automatically restored on entry to **WordSmith Tools**. If the last Tool used is **Concord**, a list of your 10 most recent search-words will be saved too.
This feature can be turned off temporarily via a menu option or permanently in **`wordsmith6.ini`** (in your Documents\wsmith6 folder).

11.36 single words v. clusters

The point of it...

Clusters are words which are found repeatedly together in each others' company, in sequence. They represent a tighter relationship than collocates, more like multi-word units or groups or phrases. (I call them *clusters* because *groups* and *phrases* already have uses in grammar and because simply being found together in software doesn't guarantee they are true multi-word *units*.) Biber 417 calls clusters, if repeated the right ways, "lexical bundles".

Language is phrasal and textual. It is not helpful to see it as a matter of selecting a word to fill a grammatical "slot" as implied by structural theories. Words keep company: the extreme example is idiom where they're bound tightly to each other, but all words have a tendency to cluster together with some others. These clustering relations may involve colligation (e.g. the relationship between **depend** and **on**), collocation 179, and semantic prosody (the tendency for **cause** to come with negative effects such as **accident, trouble,** etc.).

WordSmith Tools gives you two opportunities for identifying word clusters, in WordList 278 and Concord 175. They use different methods. Concord only processes concordance lines, while WordList processes whole texts.

How they are computed ...

Suppose your text begins like this:

Once upon a time, there was a beautiful princess. She snored. But the prince didn't.

If you've chosen 2-word clusters, the text will be split up as follows:

Once upon
upon a
a time
(note **not** "*time there*" because of the comma)
there was (etc.)
With a three-word cluster setting, it would send
Once upon a
upon a time
there was a
was a beautiful
a beautiful princess
But the prince
the prince didn't
(etc.)

That is, each n-word cluster will be stored, if it reaches n words in length, *up to a punctuation boundary*, marked by **;,.!?** (It seems reasonable to suppose that a cluster does not cross clause boundaries and these punctuation symbols help mark clause boundaries, but there is a Concord setting 163 or a WordList setting 310 for this to give you choice.)

See also: concgrams 394.

11.37 speed

networks

If you're working on a network, WordSmith will be s-l-o-w if it has to read and write results across the network. It's much faster to do your work locally on a **C:\** or **D:\** drive and then copy any useful results over to network storage later if required.

and generally

To make a word-list on 4.2 million words used to take about 20 minutes on a 1993 vintage 486-33 with 8Mb of RAM 447. The sorting procedure at the end of the processing took about 30 seconds. A 200Mz Pentium with 64MB of RAM handled over 1.7 million words per minute. On a 100Mz Pentium with 32Mb of RAM this whole process took about 3 and a half minutes, working at over a million words a minute.

When concordancing, tests on the same Pentium 100, using one 55MB text file of 9.3 million words, and a quad-speed CD-ROM drive, showed

search-word	**source**	**speed**
quickly	CD-ROM	6 million words per minute
quickly	hard disk	12 million wpm
the	CD-ROM	900,000 wpm
the	hard disk	1 million wpm
thez	CD-ROM	6 million wpm
thez	hard disk	16 million wpm

Tests using a set of text files ranging from 20K down to 4K, using *quickly* as the search-word, gave speeds of 2 million wpm rising with the longer files to 4 million wpm. Making a word list on the same set of files gave an average speed of 800,000 wpm. On the 55MB text file the speed was around 1.35 million wpm.

These data suggest that factors which slow concordancing down are, in order, word rarity (*the* was much slower than *quickly* or the non-existent *thez*), text file size (very small files of only 500 words or so (3K) will be processed about three times as slowly as big ones) and disk speed (the outdated quad speed CD-ROM being roughly half the speed of the 12ms hard disk). When Concord finds a word it has to store the concordance line and collocates and show it (so that you can decide to suspend 123 any further processing if you don't like the results or have enough already). This is a major factor slowing down the processing. Second, reading a file calls on the computer's file management system, which is quite slow in loading it, in comparison with Concord actually searching through it. Third, disk speeds are quite varied, floppy disks being much the worst for speed.

If processing seems excessively slow, close down as many programs as possible and run WordSmith Tools again. Or install more RAM. Get advice about setting Windows to run efficiently (virtual memory, disk caches, etc.) Use a large fast hard drive.

You can run other software while the programs are computing, but they will take up a lot of the processor's time. Shoot-em-up games may run too jerkily, but printing 97 a document at the same time should be fine.

11.38 status bar

The bar at the bottom of a window, which allows you to pull the whole window bigger or smaller, and which also shows a series of panels with information on the current data. The status bar can usually be revealed or hidden using a main menu option. You can right-click on the panel to bring up a popup menu offering choice between Edit, Type and Set 428.

11.39 tools for pattern-spotting

Tools are needed in almost every human endeavour, from making pottery to predicting the weather. Computer tools are useful because they enable certain actions to be performed easily, and this facility means that it becomes possible to do more complex jobs. It becomes possible to gain insights because when you can try an idea out quickly and easily, you can experiment, and from experimentation comes insight. Also, re-casting a set of data in a new form enables the human being to spot patterns.

This is ironic. The computer is an awful device for recognising patterns. It is good at addition, sorting, etc. It has a memory but it does not know or understand anything, and for a computer to recognise printed characters, never mind reading hand-writing, is a major accomplishment.

Nevertheless, the computer is a good device for helping humans to spot patterns and trends. That is why it is important to see computer tools such as these in WordSmith Tools in their true light. A tool helps you to do your job, it doesn't do your job for you.

Tool versus Product

Some software is designed as a product. A game is self-contained, so is an electronic dictionary. A word-processor, spreadsheet or database, on the other hand, is a tool because it goes beyond its own borders: you use it to achieve something which the manufacturers could not possibly anticipate. WordSmith Tools, as the name states, are not products but tools. You can use them to investigate many kinds of pattern in virtually any texts written in a good range of different languages 81.

Insight through Transformation

No, this is not a religious claim! The claim I am making is psychological. It is through changing the shape of data, reducing it and then re-casting it in a different format, that the human capacity for noticing patterns comes to the fore. The computer cannot "notice" at all (if you input 2 into a calculator and then keep asking it to double it, it will not notice what you're up to and begin to do it automatically!). Human beings are good at noticing, and particularly good at noticing visual patterns.

By transforming a text into a list, or by plotting keywords in terms of where they crop up in their source texts, the human user will tend to see a pattern. Indeed we cannot help it. Sometimes we see patterns where none was intended (e.g. in a cloud). There can be no guarantee that the pattern is "really there": it's all in the mind of the beholder.

WordSmith Tools are intended to help this process of pattern-spotting, which leads to insight. The tools in this kit are intended therefore to help you gain your own insights on your own data from your own texts.

Types of Tool

All tools take up positions on two scales: the scale of specialisation and the scale of permanence.

general-purpose ----------------- specialised

general-purpose

The spade is a digging tool which makes cutting and lifting soil easier than it otherwise would be. But it can also be used for shovelling sand or clearing snow. A sewing machine can be used to make curtains or handkerchiefs. A word-processor is general-purpose.

specialised

A thimble is dedicated to the purpose of protecting the fingers when sewing and is rarely used for anything else. An overlock device is dedicated to sewing button-holes and hems: it's better at that job than a sewing machine but its applications are specialised. A spell-checker within a word-processor is fairly specialised.

temporary ----------------- permanent

temporary

The branch a gorilla uses to pull down fruit is a temporary tool. After use it reverts to being a spare piece of tree. A plank used as a tool for smoothing concrete is similar. It doesn't get labelled as a tool though it is used as one. This kind of makeshift tool is called "quebra-galho", literally branch-breaker, in Brazilian Portuguese.

permanent

A chisel is manufactured, catalogued and sold as a permanent tool. It has a formal label in our vocabulary. Once bought, it takes up storage room and needs to be kept in good condition.

The WordSmith Tools in this kit originated from temporary tools and have become permanent. They are intended to be general-purpose tools: this is the Swiss Army knife for lexis. They won't cut your fingers but you do need to know how to use them.

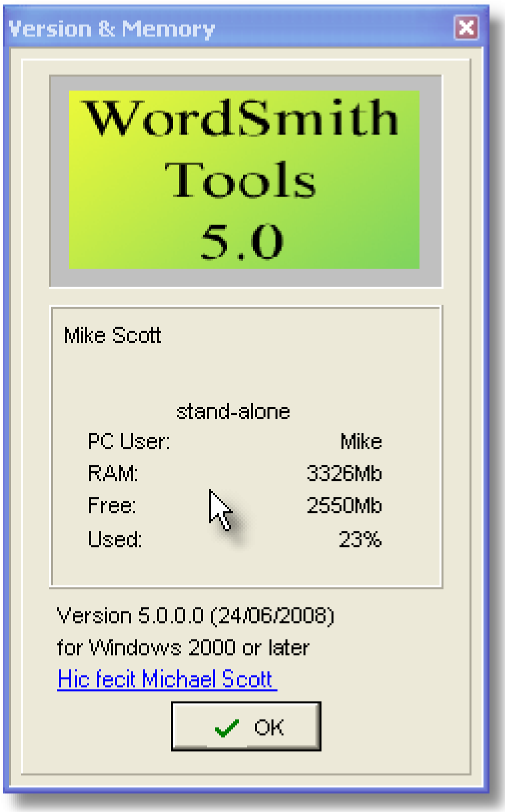
see also : Word Clouds 128, Dispersion Plots 191, Acknowledgements 416

11.40 version information

This help file is for the current version of **WordSmith Tools.**

The version of **WordSmith Tools** is displayed in the *About* option (F9) which also shows your registered name and the amount of memory 447 available. If you have a demonstration version this will be stated immediately below your name.
Check the date in this box, which will tell you how up-to-date your current version is. As suggestions are incorporated, improved versions are made available for downloading. Keep a copy of your registration code for updated versions.

You can click on the WordSmith graphic in the About box to see your current code.



Mike Scott

stand-alone

PC User: Mike
RAM: 3326Mb
Free: 2550Mb
Used: 23%

Version 5.0.0.0 (24/06/2008)
for Windows 2000 or later
Hic fecit Michael Scott

See also: 32-bit Version Differences 452, Demonstration Version 427, Contact Addresses 425.

11.40.1 Version 3 improvements

After the earlier 16-bit versions of the 1990s, WordSmith brought in lots of changes "under the hood".

- long file names
- better tag and entity 131 handling including Tag Concordancing 198
- converter for previous data
- zip file handling 453
- easier exporting of data to Microsoft Word and Excel 102
- Unicode text handling, allowing more languages 81 to be processed
- possibility of altering the data 67 as it comes in, e.g. for language-specific lemmatisation
- the old limitations of 16,000 lines of data went. (The theoretical limit for a list of data is over 134 million lines.)

See also: What's New in the current version 4, Contact Addresses 425.

11.41 zip files

Zip files are files which have been compressed in a standard way. **WordSmith** can now read and write to *.zip* files.

The point of it...

Apart from the obvious advantage of your files being considerably smaller than the originals were, the other advantage is that less disk space gets wasted like this: any text file, even a short one containing on the word "hello", will take up on your disk something like 4,000 bytes or maybe up to 32,000 depending on your system. If you have 100 short files, you would be losing many thousands of bytes of space. If you "zip" 100 short files they may fit into just 1 such space. Zip files are used a lot in Internet transmissions because of these advantages. If you have a lot of word lists to store, it will be much more efficient to store them in one .zip file.

The "cost" of zipping is a) the very small amount of time this takes, b) the resulting .zip file can only be read by software which understands the standard format. There are numerous zip programs on the market, including *PKZip*™ and *Winzip*™. If you zip up a word list, these programs can unzip it but won't be able to do anything with the finished list. **WordSmith** can first unzip it and then show it to you.

How to do it...

Where you see an option to create a zip file, this can be checked, and the results will be stored where you choose but in zipped form with the *.zip* ending.

If you choose to open a zipped word list, concordance, text file, etc. and it contains more than one file within it, you will get a chance to decide which file(s) within it to open up. Otherwise the process will happen in the background and will not affect your normal **WordSmith** processing.

Troubleshooting

Section XII

12 Troubleshooting

12.1 list of FAQs

See also: logging 31.

These are the Frequently Asked Questions.

There's a much longer list of explanations under Error Messages 461.

Can't process apostrophes 455
Is this Russian, Greek or English? strange symbols in display 456
It crashed 456
It doesn't even start! 458
It takes ages! 458
Keys don't respond 457
Line beyond demo limit 456
Mismatch between Concord and WordList results 456
No tags visible in concordance 455
Printing problem 458
Text is unreadable because of the colours 457
Too much or too little space between columns 455
Wordlist out of order 459
Won't slice pineapples 458

12.2 apostrophes not found

Apostrophes not processed

If your original text files were saved using Microsoft Word™, you may find **Concord** can't find apostrophes or quotation marks in them! This is because Word can be set to produce "smart" symbols. The ordinary apostrophe or inverted comma in this case will be replaced by a curly one, curling left or right depending on its position on the left or right of a word. These smart symbols are not the same as straight apostrophes or double quote symbols.

Solution: select the symbol in the character set in the Controller, then paste when entering your search word 159, or else replace them in your text files using Text Converter 355.
See also: settings 113

12.3 column spacing

column spacing is wrong

You can alter this by clicking on the layout 87 button.

12.4 Concord tags problem

no tags visible in concordance

If you can't see any tags after asking for *Nearest Tag* in **Concord**, it is probably because the Tags to Ignore 132 has the same format. For example, if *Text to Ignore* has **<*>**, any tags such as **<title>**, **<quote>**, etc. will be cut out of the concordance unless you specify them in a tag file 141.
Solution: specify the tag file and run the concordance again.

12.5 Concord/WordList mismatch

Concord/WordList mismatch

If **WordList** finds a certain number of occurrences of a (word list) cluster 448 but **Concord** finds a different number, this is because the procedures are different. WordList proceeds word by word, ignoring punctuation (except for hyphens and apostrophes). When **Concord** searches for a (concordance) cluster 175 it will (by default) take punctuation into account: you can change that in the settings 222 if you wish.

12.6 crashed

it crashed!

Solution: quit **WordSmith Tools** and enter again. If that fails, quit Windows and try again.
Or try logging 32. The idea of Logging is to find out what is causing a crash. It is designed for when WS gets only part of the way through some process. As it proceeds, it keeps adding messages to the log about what it has found & done. When it crashes, it can't add any more messages! So if you examine the log you can see where it was up to. At that point, you may see a text file name that it opened up. Examine that text, you might be able to see something strange about it, eg. it has got corrupted.

12.7 demo limit

demo limit reached

You may have just downloaded, but you haven't yet supplied your registration details. To do this, go to the main WordSmith Tools window, and choose *Settings* | *Register* in the menu.

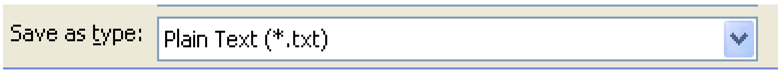
If you haven't got the registration code, contact Lexical Analysis Software (sales@lexically.net). The ***only*** difference between a demonstration version 427 and a full version is: with the latter you can see or print all the data, with the former you'll be able to see only about 25 lines of output.

12.8 funny symbols

weird symbols

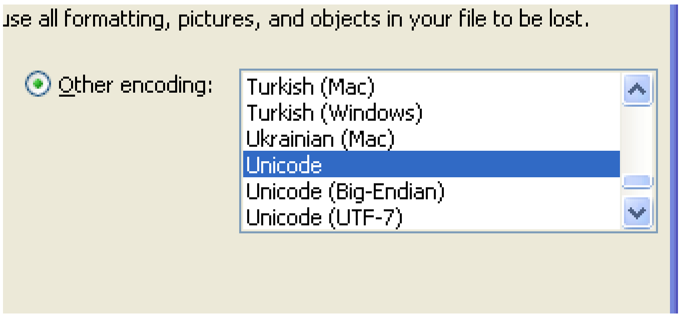
funny symbols when using WordSmith Tools

1. Check your text files. Look at them in **Notepad**. Do they contain lots of strange symbols? These may be hidden codes used by your usual word-processor. Solution: open them in your usual word-processor and *Save As*, with a new name, in plain text format, sometimes called "Text Only" or **.txt**. In Word 2003 the option looks like this:



Save as type: Plain Text (*.txt)

and then choose Unicode:



use all formatting, pictures, and objects in your file to be lost.

● Other encoding:

- Turkish (Mac)
- Turkish (Windows)
- Ukrainian (Mac)
- Unicode
- Unicode (Big-Endian)
- Unicode (UTF-7)

2. *Choose Texts*, select the text file(s), right-click and *View*. Does it contain strange symbols?
3. Use Text Converter 365 to clean up and convert and your text files to Unicode.

Greek, Russian, etc.

4. If the text is in Russian, Greek, etc. you will need an appropriate font, obtainable from your Windows cd or via the Microsoft website.
5. If you have several lists open which use *different* character sets, and you change Font 78 or Text Characteristics 124, the lists will all be updated to show the current font and character set, unless you first minimize any window which would be affected.

funny symbols when reading WordSmith data in another application

WordSmith Tools can Save 101 or Save As and Saves as text 102 by printing 97 to a file. "Save" and "Save As" will store the file in a format for re-use by **WordSmith**. This format is not suitable for reading into a word processor. The idea is simply for you to store your work so that you can return to it another day.

"Save as Text", on the other hand, means saving as plain text, by "printing" to a file. This function is useful if you don't want to print to paper from **WordSmith** but instead take the data into a spreadsheet, or word processor such as Microsoft Word. It is usually quicker to copy the selected text into the clipboard 422.

12.9 illegible colours

text unreadable because of colours

Solution: in *Settings*, choose *Colours*. You can now set the colours which suit your computer monitor. Monochrome settings are available.

12.10 keys don't respond

Keys don't respond

If a key press does nothing, it is probably because the wrong window, or the wrong column in the window, has the focus. As you know, Windows is designed to let users open up a number of programs at once on the same screen, so each window will respond to different key-press combinations. You can see which window has the focus because its caption is coloured differently from all the others. The solution is to click within the appropriate window/column, then press the key you wanted.

12.11 pineapple-slicing

won't slice a pineapple

"*Propose to any Englishman any principle, or any instrument, however admirable, and you will observe that the whole effort of the English mind is directed to find a difficulty, a defect, or an impossibility in it. If you speak to him of a machine for peeling a potato, he will pronounce it impossible: if you peel a potato with it before his eyes, he will declare it useless, because it will not slice a pineapple.*" Charles Babbage, 1852.
(Babbage was the father of computing, a 19th Century inventor who designed a mechanical computer, a mass of brass levers and cog-wheels. But in order to make it, he needed much greater accuracy than existing technology provided, and had all sorts of problems, technical and financial. He solved most of the former but not the latter, and died before he was able to see his Difference Engine working. The proof that his design was correct was shown later, when working versions were made. The difficulties he encountered in getting support from his government weren't exclusively English.)

12.12 printer didn't print

printing problem

If your printing comes out with one or more columns printed OK but others blank, you may have pulled your columns too wide for the paper. WordSmith uses information about your printer's defaults to compute what will and will not fit on the current paper. If you can change the printer settings to landscape that will give more space.

12.13 too slow

It takes ages

If you're processing a lot of text and you have an ancient PC with little memory and a hard disk that Noah bought from a man in the market for a rainy day, it might take ages. You'll hear a lot of clicks coming from the hard disk when memory 447 is low. Solution: get a faster computer, by installing more memory which makes a *big* difference), by defragmenting your hard drive, by using a disk cache, or by adjusting virtual memory settings. If you're running **WordSmith Tools** on a network, check with the network administrator whether performance is significantly degraded because of network access.

Solution 2: quit all programs you don't need. That can restore a lot of system memory.

Solution 3: quit Windows and start again. That can restore a lot of system memory.

Solution 4: save and read from the local hard disk (C: or D:), not the network.

12.14 won't start

it doesn't even start

Yikes!

12.15 word list out of order

word-list out of order

Words are sorted according to Microsoft routines which depend on the language. If you process Spanish but leave the Language settings to "English", you will get results which are not in correct Spanish order, (e.g. **`LL`** will come just before **`LM`**).
Solution: choose your language 81 and re-compute the word-list.

Error Messages

Section XIII

13 Error Messages

13.1 list of error messages

List of Error Messages

See also: Troubleshooting 455.

13.2 .ini file not found

.ini file not found

On starting up, **WordSmith** looks for the **`wordsmith6.ini`** file which holds your current defaults 113. If you've removed or renamed it, restore it. This file should be in a sub-folder of your Documents folder called \wsmith6.

13.3 administrator rights

administrator rights

If you see this error message it's because you need Administrator rights to register WordSmith.

Try searching for "Run as Administrator" or this link.

13.4 base list error

base list error

WordSmith is trying to access an word or concordance line above or below the top or bottom of the data computed. This is a bug.

13.5 can only save words as ASCII

Can only save WORDS as Plain Text

WordSmith Tools can't save graphics as a text file. If you get this error message, you can only save this type of data by copying to the clipboard 422 and pasting it into your word-processor.

13.6 can't call other tool

Can't call other Tool

Inter-Tool communication has got disrupted. Save 211 your work, first. Then, if necessary, close down WordSmith Tools altogether, then start the main `wordsmith6.exe` program again.

13.7 can't make folder as that's an existing filename

Can't make folder as that's an existing filename

If you already have a *file* called C:\TEMP\FRED, you can't make a *sub-folder* of C:\TEMP called FRED. Choose a new name.

13.8 can't compute key words as languages differ

Can't compute key words as languages differ

Key words can only be computed if both the text file and the reference corpus are in the same primary language. You can compute KWs using 2 different varieties of English or 2 different varieties of Spanish, but not between English and French.

13.9 can't merge list with itself!

Can't merge list with itself

You can only merge 1 word list or key word database with 1 other at a time. Select (by clicking while holding down the Control key) 2 file-names in the list of files.

13.10 can't read file

Can't read file

If this happens when starting up WordSmith Tools, there is probably a component file missing. One example is sayings.txt, which holds sayings that appear in the main Controller 4 window. If you've deleted it, I suggest you use notepad to start a new sayings.txt and put one blank line in it.

If you get this message at another time, something has gone wrong with a disk reading operation. The file you're trying to read in may be corrupted. This happens easily if you often handle very large files. See your Windows manual for help on fragmentation.

13.11 character set reset to <x> to suit <language>

Character set reset to <x> to suit <language>

Prior to version 2.00.07, WordSmith Tools handled fewer character sets 419 and languages 81 than it does now. Accordingly, data saved in the format used before that version may not "know" what language it was based on. If you get this message when opening up an old WordSmith data file, it's because WordSmith doesn't know what language it derived from. Through gross linguistic imperialism, it will by default assume that the language is English!

If the data are okay, just click the save button so that next time it will "know" which language it's based on. If not, reset the language to the one you want in the Controller 4, Language Settings | Text, then re-save the list.

13.12 concordance file is faulty

Concordance file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **`.CNC, .LST`**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **Concord**.

13.13 concordance stop list file not found

Concordance stop list file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.14 confirmation messages: okay to re-read

Okay to re-read?

A confirmation message. To proceed, **Viewer & Aligner** will now re-read the disk file. This will affect any alterations you've already made to the display. You may wish to save first and then try again later.
Also, Viewer & Aligner will try to read the whole text file. If you have a very big file on a slow CD-ROM drive, this will take some time.

13.15 conversion file not found

Conversion file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.16 destination folder not found

Destination folder not found

WordSmith couldn't find that folder; perhaps it's mis-spelt.

13.17 disk problem -- file not saved

Disk problem: File not saved

Something has gone wrong with a disk writing operation. Perhaps there's not enough room on the drive. If so, delete some files on that drive.

13.18 dispersions go with concordances

Dispersions go with concordances

They can't be saved 211 separately.

13.19 drive not valid

Drive not valid

WordSmith is unable to access this drive. This could happen if you attempt to access a disk drive which doesn't exist, e.g. drive P: where your drives include A:, C:, D: and E:.

13.20 failed to access Internet

Failed to access Internet

This function relies on a) your having an Internet browser on your computer, b) your system "associating" an Internet URL ending **.htm** with that browser.

13.21 failed to create new folder name

Failed to create new folder or file-name

A folder and a file cannot have the same name. If you already have a *file* called **C:\TEMP\FRED**, you can't make a *sub-folder* of **C:\TEMP** called **FRED**. Choose a new name.
Or you don't have rights to create files in that folder. Or something went wrong while WordSmith was trying to write a file, for example the disk was full up.

13.22 failed to read file

Failed to Read

This may have happened
a) because you included a text file which happens to be empty (zero size), or
b) because your disk filing system has got screwed up, which is especially likely to occur if you often use large files in your word processor (in which do a disk cleanup) or
c) because you tried to use the wrong kind of file for the job (for example the KeyWords procedure won't work if you choose text files as your word-lists).

13.23 failed to save file

Failed to Save

Maybe because you had the same file open in another program or another instance of the Tool you're running. If so, close it and try again.
Or because the folder you're saving to is a read-only folder on a network, or because the disk is full, or because your disk filing system has got screwed up. This last problem is quite common, actually, and is especially likely to occur if you often use large files in your word processor. In that case run *Programs* | *Accessories* | *System Tools* | *Disk Defragmenter*.
If you're working on a network, you will be able to save 211 on certain drives and folders but not others; the solution is to try again on a memory stick or a hard disk drive which you do have the right to save to.

13.24 file access denied

File Access Denied

Maybe the file you want is already in use by another program. You'll find most word-processors label any text files open in them as "in use", and won't let other programs access them even just to read them. Close the text file down in your word processor.

13.25 file contains none of the tags specified

File contains none of the tags specified

You specified tags, but none of them were found.

13.26 file has "holes"

File has "holes"

Text files are supposed to contain only characters, punctuation, numbers, etc. without any unrecognised ones such as character(0). The problem could have arisen because it was transferred from one system to another, part of the disk is corrupted, or else maybe the file contains unrecognised graphics (or else it is not a plain text file but e.g. a Word document) 473.

You can solve this problem by converting the text using the Text Converter. If it is a plain text with holes these will be replaced by spaces. You can find texts with holes using the File Utilities.

13.27 file not found

File not found

This message, like Original Text not found 474, can appear when WordSmith needs to access the original source text used when a list was created, but cannot find it. Have you deleted or moved it? If the file is still available, you may be able to edit the file names in the file name window (≣) of this list.

Or the message may come after you've supplied the file name yourself. You may have mis-typed it. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.28 filenames must differ!

Filenames must differ

You can't compare a file with itself.

13.29 folder is read-only

folder is read-only

For some purposes, WordSmith needs to save files e.g. lists of results you have made so that you can get at recent files again. To do this it needs a place where your network or operating system lets you save. Usually \wsmith6 is fine, but in some institutional settings the drive or folder may be "read-only". If you see this message, choose Folder Settings and select there a folder where you can write as well as read.

13.30 for use on X machine only

For use on pc named XXX only

The software was registered for use on another PC. If you get this message, please re-install as appropriate.

13.31 form incomplete

Form incomplete

You tried to close a form where one or more of the blanks needed to be filled in before **WordSmith** could proceed.

13.32 full drive & folder name needed

Full drive:\folder name needed

When typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.33 function not working properly yet

function not working properly yet

This is a function under development, still not fully implemented.

13.34 invalid concordance file

Invalid Concordance file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **`.CNC, .LST`**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **Concord**.

13.35 invalid file name

Invalid file name

File names may not contain spaces or certain symbols such as **?** and *****.

13.36 invalid KeyWords database file

Invalid Keywords Database file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **`.KWS, .KDB`**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .KDB file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced for a database by the current version of **KeyWords**.

13.37 invalid KeyWords calculation

Invalid Keywords calculation

For KeyWords to calculate the key-words in a text file by comparing it with a reference corpus, both must be in the same language, both must be sorted in the same way (alphabetical order, ascending) and they should both be in the same format (Unicode or single-byte). If you see this message you are trying to compute KWs without meeting these criteria. Solution: open each word-list and check to see it is OK and that it is sorted alphabetically in the same way (in the Alphabetical view, click the top bar to re-sort in ascending alphabetical order), then save it. Check they have both been made with the same language & format settings and if necessary re-compute one or both of them.

13.38 invalid WordList comparison file

Invalid Wordlist Comparison file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **`.LST, .CNC`**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced as a comparison file by **WordList**.

13.39 invalid WordList file

Invalid Wordlist file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **`.LST, .CNC`**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .LST file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **WordList**.

13.40 joining limit reached

Joining limit reached: join & try again

Only a certain number of words can be lemmatised 270 in one operation. If you reach the limit and get this message,
1. lemmatise by pressing F4,

2. place the highlight on the head entry again
3. press F5 and carry on lemmatising by pressing F5 on each entry you wish to attach to the head entry
4. when you've done, press F4 to join them up.

13.41 KeyWords database file is faulty

Keywords Database file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .**KDB, .KWS**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .KDB file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced for a database of keywords, by the current version of **KeyWords**.

13.42 KeyWords file is faulty

Key words file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **.KWS, .KDB**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .KWS file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **KeyWords**.

13.43 limit of file-based search-words reached

Limit of search-words reached

No more than 15 search-words can be processed at once, unless you use a file of search words 161 to tell Concord to do them in a batch, where the limit is 500.

13.44 links between Tools disrupted

Links between Tools disrupted

WordSmith Tools Controller 4 or an individual Tool has tried to call another Tool and failed. There may have been a fault in another program you're running or a shortage of memory. As inter-tool communication links 438 are vital in this suite, you should exit WordSmith and re-enter.

13.45 match list details not specified

Match list details not specified

You pressed the Match List 92 button but then failed to choose a valid match list file or else to type in a template for filtering. Try again.

13.46 must be a number

Must be a number

You typed in something other than a number. Be especially careful with lower-case **L** and **1**, and **O** (the letter) instead of **0** (the number).

13.47 mutual information incompatible

Mutual information list is incompatible

A mutual information list derives from an index file, and knows which index file it derives from when computed. Normally when it opens up, it opens up the corresponding index file too. If that index If the file is not found on your PC or has been renamed, you will see this message. The mutual information can still be accessed but a) what you see in terms of Frequency and Alphabetical lists refers to a different index file, and b) it will not be possible to get concordances directly from the listing.

13.48 network registration used elsewhere

Network registration running elsewhere or vice-versa

The site licence registration for use on a network is not valid for use on a stand-alone pc, and vice-versa. If you get this message, please re-install as appropriate.

13.49 no access to text file - in use elsewhere?

No access to text file: in use elsewhere?

The file cannot be accessed. Perhaps another application is using it. If so, close down the file in that other application and try again.

13.50 no associates found

No associates found

Alter settings (*Settings* | *Min & Max Frequencies*) and try again.

13.51 no clumps identified

No clumps identified

Alter settings and try again.

13.52 no clusters found

No clusters found

Alter the settings (*Settings* | *Clusters*) and try again. There were too few concordance lines to find the minimum number needed, or the cluster length was too great.

13.53 no collocates found

No collocates found

In the Controller 4, alter the settings (Concord settings | Min. Frequency) and try again. There were too few concordance lines to find the minimum number needed.

13.54 no concordance entries

No concordance entries found

If you got no concordance entries, either a) there really aren't any in your text(s), b) there's a problem with the specification of what you're seeking, or c) there's a problem with the text selection. Check how you've spelt the search-word and context word. If you're using accented text 419, check the format of your texts. If you're using a search-word file 161, ensure this was prepared using a plain Windows word-processor such as Notepad.

Have you specified any wildcards 159 (* and ?) accurately? If you are looking for a question-mark, you may have put **"?"** correctly but remember that question-marks usually come at the ends of words, so you will need ***"?"**.

Tip

Bung in an asterisk 159 or two. You're more likely to find book* than book.

13.55 no concordance stop list words

No concordance stop list words

13.56 no deleted lines to zap

No deleted lines to Zap

You pressed Ctrl+Z but hadn't any deleted lines to zap 129. No harm done.

13.57 no entries in KeyWords database

No entries in Keywords Database

Alter settings and try again.

13.58 no fonts available

no fonts available for language

The operating system does not have a font which can show the characters for that language. You need to find and install a font.

13.59 no key words found

No Key Words found

Alter settings and try again. The minimum frequency is set too high and/or the p value 235 too small for any key words to be detected. For very short texts a minimum frequency of 2 may be needed.

13.60 no key words to plot

No key words to plot

Had you deleted them all?

13.61 no KeyWords stop list words

No keyword stop list words

WordSmith either failed to read your stop-list file or it was empty.

13.62 no lemma list words

No lemma match list words

WordSmith either failed to read your lemma list file or it was empty.

13.63 no match list words

No match list words

WordSmith either failed to read your match list 92 file, or it was empty, or you forgot to check the action to be taken (one option is *None*). Or you tried to match up using a list of words, or a template, when the current column has only numbers. Or else there really aren't any like those you specified!

13.64 no room for computed variable

No room for computed variable

There isn't enough space for the variable you're trying to compute.

13.65 no statistics available

No statistics available

Some types of word list created by **WordSmith Tools**, e.g. a word list of a key words database have words in alphabetical and frequency order but no statistics on the original text files. You cannot therefore call the statistics up in **WordList**. You might also see this message if the statistics file you're trying to call up is corrupted.

13.66 no stop list words

No stop list words

WordSmith either failed to read your stop-list file or it was empty.

13.67 no such file(s) found

No such file(s) found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.68 no tag list words

No tag list words

WordSmith either failed to read your tag file or it was empty.

13.69 no word lists selected

No word lists selected

For **WordSmith** to know which word lists to compare, you need to select them, by clicking on one in each folder. If you've changed your mind, press Cancel.

13.70 not a valid number

Not a valid number

Either you've just typed in, or else **WordSmith Tools** has just attempted to read (e.g. from **`wordsmith6.ini`**, the defaults 113 file), something which is expected to be a number but wasn't. Computers will not see the capital **O** as equivalent to the number **0**. Or else there is a number but accompanied by some other letters or symbols, e.g. ***£30***. If this happens when **WordSmith** is starting up, check out the **`wordsmith6.ini`** file for mistakes.

13.71 not a WordSmith file

The file you are trying to open is not a WordSmith Tools file. WordSmith makes files containing your results, files whose names end in **`.LST, .CNC, .KWS`**, etc. These are in WordSmith's own format and cannot be opened up by Microsoft Word -- likewise a plain text file or a **`Word .doc`** 444 cannot usually be read in by WordSmith as a data file, but only as a text file for processing.

See also: Converting Data from Previous Versions 323

13.72 not a current WordSmith file

Not a Current WordSmith File

The file you are trying to open was made using WordSmith but either

- it's a file made using version 1-3
  or
- it's a file made with the beta version of WordSmith and the format has had to change (sorry!)

If the former, you may be able to convert it using the Converter 323.

13.73 nothing activated

Nothing activated

Some forms have choices labelled "Activated" which you can switch on and off. If they are unchecked, you can still see what they would be but **WordSmith** will ignore them.

13.74 Only X% of words found in reference corpus

Only X% of words found in reference corpus

When WordSmith computes key words it checks to see that most of the words in your small word-list are found in the reference corpus, as would be expected. If less than 50% are found, you will get this warning. That is a bit unusual, and is supplied as a warning that for example there might be something strange about one of your two texts. If you know there is nothing strange, then you could ignore the message.

If you are processing clusters you are much more likely to see this warning, however, as the chance of 3-word strings matching in the two lists is less than that of single words matching.

It is up to you to decide whether there is some error in what you are doing or it is OK for many of your smaller word list's words/clusters not to be found in the reference corpus word list.

It might not be so unusual if your reference corpus was very small. But if it is indeed very small, the whole procedure is not very reliable. WordSmith simply looks at the frequencies of each word form and uses basic statistics to compute how greatly they differ in frequency. Basic statistics rely on a notion of what can be expected. If the reference corpus is incredibly small, WordSmith's computation of what is to be expected isn't really very reliable. As a dumb example if you met three citizens of a country you have never visited, and all looked fat, you might suppose the people of that country to be fat in general, but the sample size is not reliable for such an expectation. The KW procedure isn't really proof of anything, incidentally. Words don't occur in texts at all randomly and all ordinary basic statistics can do in my opinion is give us food for thought. So a KW listing isn't proof of anything but it may well give good ideas as to what may prove interesting avenues for research.

13.75 original text file needed but not found

Original text file(s) needed but not found

To proceed, **WordSmith** needed to find the original text file 113 which the list was based on. But it has been moved or renamed.
Or if on a network, your network connection is not mapped, or the network is down ...or else the right disk or CD-ROM is not in the drive!

13.76 printer needed

WordSmith needs a printer driver to be installed, even if you never actually print anything. You don't need to buy a printer or to switch a printer on, but the Print Preview 97 function in Concord, WordList, KeyWords etc. does need to know what sort of paper size you would print to. If you get a message complaining that no printer has been installed, choose Start | Settings | Printers & Faxes and install a default printer (any printer will do) in Windows.

13.77 registration code in wrong format

Registration code unexpectedly short

PASTE the registration supplied into the box; only paste into the Name or Other Details boxes the details supplied.

If you see this message on registering you may have a registration for a previous major version. If so, contact sales at lexically dot net with your original purchase details and you will be entitled to a

50% discount on the current version.

13.78 registration is not correct

Registration is not correct

It doesn't match up with what's required for a full updated version! The old registration code in earlier versions is no longer in use. WordSmith will still run but in Demonstration Version 427 mode.

13.79 short of memory

Short of Memory!

An operation could not be completed because of shortage of RAM 447.

13.80 source folder file(s) not found

Source Folder file(s) not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the filename itself.

13.81 stop list file not found

Stop list file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.82 stop list file not read

Stop list file not read

Something has gone wrong with a disk reading operation. The file you're trying to read in may be corrupted. This happens easily if you often handle very large files, especially if it's a long time since you last ran *Scandisk* to check whether any clusters in your files have got lost. See your DOS or Windows manual for help on fragmentation.

13.83 tag file not found

Tag File not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.84 tag file not read

Tag list file not read

Something has gone wrong with a disk reading operation. The file you're trying to read in may be corrupted. This happens easily if you often handle very large files. See your Windows manual for help on fragmentation.

13.85 this function is not yet ready

This function is not yet ready!

Temporary message, for functions which are still being tested.

13.86 this is a demo version

This is a demo version

You will probably want to upgrade 427 to the full version.

13.87 this program needs Windows XP or greater

This program needs Windows XP or better

From version 4.0, this program has required operating systems for this millennium.

13.88 to stop getting this message ...

Get an update. This is "annoyware" for the demonstration version 427.

13.89 too many requests to ignore matching clumps

The limit is 50. Do any remaining joining manually.

13.90 too many sentences

The limit is 8,000. Do the task in pieces.

13.91 truncating at xx words -- tag list file has more

The tag list file has more entries than the current limit. Or else it isn't a tag list file at all!

13.92 two files needed

You need to select 2 files for this procedure. Select (by clicking while holding down the Control key) 2 file-names in the list of files.

13.93 unable to merge Keywords Databases

Perhaps there wasn't enough RAM 447 to carry out the merge.

13.94 why did my search fail?

The standard search function (F12 or 🔍) for a list of data operates on the currently highlighted column. If you want to search within data from another column, click in that column first.

By default, a search is "whole word". Use * at either end of the word or number you're searching for if you want to find it, e.g. in any data consisting of more than one word. (The advantage of the asterisk system is that it allows you to specify either a prefix or a suffix or both, unlike the standard

Windows search "whole word" option.)

13.95 word list file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **`.LST, .KWS`**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **WordList**.

13.96 word list file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.97 WordList comparison file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **`.LST, .KWS`**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced as a comparison file by **WordList**.

13.98 WordSmith Tools already running

Don't try to start **WordSmith Tools** again if it's already running. Just Alt-tab back to the instance which is running. (You can, however, have several copies of each tool running at once.)

13.99 WordSmith Tools expired

Message for limited period users only. Your version of WordSmith Tools has passed its validity and is now in demo 427 mode. Download another from the Internet 425.

13.100WordSmith version mis-match

Since the various Tools are linked 438 to each other, it is important to ensure that the component files are compatible with each other. If you get this message it is because one or more components is dated differently from the others.

Solution: download those you need from one of the contact websites 425.

13.101XX days left

Message for limited period users only. At the end of this time WordSmith will revert to demo 427 mode.

Index

WordSmith Tools Manual