

Oxford WordSmith Tools

Version 4.0

© 2004-2005 Mike Scott

Oxford WordSmith Tools

version 4.0

by Mike Scott

© 2004-2005 Mike Scott

WordSmith Tools

© 2004-2005 Mike Scott

All rights reserved. No parts of this work may be reproduced in any form or by any means - graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems - without the written permission of the publisher.

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

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

Printed: October 2005

Publisher

Oxford University Press

Special thanks to:

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

Table of Contents

Foreword	1
Part I WordSmith Tools	3
Part II Overview	5
1 What's new in version 4	5
2 Controller	5
3 Concord	5
4 KeyWords	6
5 WordList	6
6 Utilities	6
Choose Languages	6
File Utilities	7
Minimal Pairs	7
Splitter	7
Text Converter	7
Version Checker	8
Viewer	9
Webgetter	9
Part III Getting Started	12
1 getting started with Concord	12
2 getting started with KeyWords	13
3 getting started with WordList	14
Part IV Installation and Updating	16
1 installing WordSmith Tools	16
2 network defaults	16
3 version checking	17
Part V Controller	19
1 adjust settings	19
2 accents	19
3 add notes	19
4 batch processing	19
5 batch folders	22
6 change font	22
7 choose favourite texts	23
8 choose language	23
9 choose texts	24

10	choosing files from standard dialogue box	26
11	colours	27
12	column totals	29
13	compute new column of data	30
14	count data frequencies	30
15	copy your results	32
16	custom processing	33
17	customising popup menus	36
18	editing a list of data	38
19	find relevant files	39
20	class or session instructions	40
21	layout & format	40
22	match words in list	42
23	print and print preview	44
24	never used WordSmith before	45
25	previous lists	45
26	quit WordSmith	45
27	reduce data to n entries	45
28	save as text	46
29	save defaults	48
30	save results	49
31	search & replace	50
32	search by typing	50
33	search for word or part of word	50
34	see filenames	50
35	stop lists	51
36	suspend processing	51
37	text and languages	52
38	window management	54
39	zap unwanted lines	54
Part VI Tags and Markup		57
1	overview	57
2	tag-types	57
3	handling tags	58
4	multimedia tags	59
5	tags as selectors	60
6	only if containing... ..	61
7	selecting within texts	62
8	making a tag file	62

9	start and end of text segments	64
Part VII Concord		67
1	purpose	67
2	what is a concordance	67
3	blanking	67
4	categories	68
5	collocate horizons	69
6	collocate settings	69
7	collocates display	70
8	collocation relationship	71
9	collocation	72
10	Concord: clusters	73
11	Concord: dispersion	74
12	Concord: saving and printing	76
13	Concord: viewing options	76
14	Concord: handling sounds & video	76
15	Concord: what you see and do	77
16	concordance settings	79
17	concordancing on tags	81
18	context word	82
19	editing concordances	82
20	file-based search-words	83
21	nearest tag	84
22	patterns	86
23	re-sorting	87
24	re-sorting: collocates	88
25	re-sorting: dispersion plot	88
26	text segments in Concord	88
27	search word syntax	89
28	WordSmith controller: Concord: settings	90
Part VIII KeyWords		94
1	purpose	94
2	Two word-list analysis	94
3	associate definition	94
4	associates	95
5	choosing files	95
6	clumps	95
7	KeyWords clusters	96

8 concordance	97
9 creating a database	97
10 example of key words	97
11 key key-word definition	98
12 key-ness definition	98
13 KeyWords database	99
14 KeyWords: advice	99
15 KeyWords: calculation	99
16 KeyWords: links	100
17 make a word list from keywords data	101
18 p value	101
19 plot calculation	101
20 plot display	101
21 regrouping clumps	103
22 re-sorting: KeyWords	103
23 the key words screen	103
24 WordSmith controller: KeyWords settings	104
Part IX WordList	107
1 purpose	107
2 auto-joining lemmas	107
3 choosing lemma file	108
4 comparing wordlists	109
5 merging wordlists	109
6 comparison display	110
7 consistency analysis (detailed)	111
8 consistency analysis (simple)	112
9 lemmas	113
10 index lists: uses	115
11 index lists: viewing	115
12 making a WordList Index	116
13 index clusters	117
14 menu search	120
15 mutual information scores	120
16 mutual information: computing	122
17 mutual information display	122
18 re-sorting: consistency lists	124
19 statistics	125
20 import words from text list	125
21 type/token ratios	127

22	case sensitivity	128
23	minimum & maximum settings	128
24	sort order	129
25	WordList display	130
26	WordSmith controller: WordList settings	132
Part X Utility Programs		135
1	Convert Data from Previous Versions	135
	Convert Data from Previous Versions	135
2	WebGetter	135
	overview	135
	settings	135
	display	137
	limitations	138
3	Languages Chooser	138
	Overview	138
	Language	139
	Font	141
	Sort Order	141
	Other Languages	142
	saving your choices	142
4	Minimal Pairs	142
	aim	142
	requirements	142
	choosing your files	143
	output	143
	rules and settings	144
	running the program	144
5	File Utilities	145
	index	145
	Splitter	145
	aim of Splitter.....	145
	Splitter: filenames.....	146
	Splitter: wildcards.....	146
	join text files	146
	compare two files	148
	file chunker	148
	find duplicates	149
	rename	149
6	Text Converter	150
	purpose	150
	Text Converter: settings	151
	Text Converter: syntax	152
	Text Converter: move if	154
	Text Converter conversion file	154
	Text Converter: sample conversion file	155
Part XI Viewer and Aligner		157
1	purpose	157

2	aligning with Viewer	157
3	aligning and moving	157
4	editing	158
5	numbering sentences & paragraphs	158
6	options	158
7	sentence joining and splitting	159
8	settings	159
9	technical aspects	159
10	translation mis-matches	160
11	troubleshooting	160
12	unusual sentences	160
Part XII Reference		163
1	32-bit version	163
2	acknowledgements	163
3	bibliography	163
4	bugs	164
5	Character Sets	165
	overview	165
	accents & symbols	166
	ansi and ascii	166
	DOS	166
	Windows	167
	Unicode	167
6	clipboard	167
7	contact addresses	170
8	date format	170
9	Definitions	170
	definitions	170
	word separators	171
10	demonstration version	171
11	edit v. type-in mode	172
12	finding source texts	172
13	folders\;directories	172
14	HistoryList	173
15	HTML, SGML and XML	173
16	hyphens	173
17	Internet	174
18	limitations	174
19	tool-specific limitations	174
20	links between tools	175
21	list of menu and button options	176

22	keyboard shortcuts	178
23	long file names	179
24	machine requirements	179
25	manual for WordSmith Tools	179
26	numbers	180
27	plot dispersion value	180
28	printer settings	181
29	RAM availability	182
30	reference corpus	182
31	restore last file	182
32	single words v. clusters	183
33	speed	183
34	status bar	184
35	tools for pattern-spotting	184
36	version information	185
37	zip files	186
Part XIII Troubleshooting		188
1	list of FAQs	188
2	apostrophes not found	188
3	column spacing	188
4	Concord tags problem	188
5	Concord/WordList mismatch	189
6	crashed	189
7	demo limit	189
8	funny symbols	189
9	illegible colours	190
10	keys don't respond	190
11	pineapple-slicing	190
12	printer didn't print	190
13	too slow	190
14	won't start	191
15	word list out of order	191
Part XIV Error Messages		193
1	list of error messages	193
2	.ini file not found	194
3	base list error	194
4	can only save words as ASCII	195
5	can't call other tool	195

6	can't make folder as that's an existing filename	195
7	can't compute key words as languages differ	195
8	can't merge list with itself!	195
9	can't read file	195
10	character set reset to <x> to suit <language>	196
11	concordance file is faulty	196
12	concordance stop list file not found	196
13	confirmation messages: okay to re-read	196
14	conversion file not found	196
15	destination folder not found	196
16	disk problem -- file not saved	196
17	dispersions go with concordances	197
18	drive not valid	197
19	failed to access Internet	197
20	failed to create new folder name	197
21	failed to read file	197
22	failed to save file	197
23	file access denied	197
24	file contains none of the tags specified	198
25	file has "holes"	198
26	file not found	198
27	filenames must differ!	198
28	for use on X machine only	198
29	form incomplete	198
30	full drive & folder name needed	199
31	function not working properly yet	199
32	invalid concordance file	199
33	invalid file name	199
34	invalid KeyWords database file	199
35	invalid KeyWords calculation	199
36	invalid WordList comparison file	199
37	invalid WordList file	200
38	joining limit reached	200
39	KeyWords database file is faulty	200
40	KeyWords file is faulty	200
41	limit of file-based search-words reached	200
42	links between Tools disrupted	200
43	match list details not specified	201
44	must be a number	201

45	mutual information incompatible	201
46	network registration used elsewhere	201
47	no access to text file - in use elsewhere?	201
48	no associates found	201
49	no clumps identified	201
50	no clusters found	201
51	no collocates found	202
52	no concordance entries	202
53	no concordance stop list words	202
54	no deleted lines to zap	202
55	no entries in KeyWords database	202
56	no key words found	202
57	no key words to plot	202
58	no KeyWords stop list words	203
59	no lemma list words	203
60	no match list words	203
61	no room for computed variable	203
62	no statistics available	203
63	no stop list words	203
64	no such file(s) found	203
65	no tag list words	203
66	no word lists selected	204
67	not a valid number	204
68	not a WordSmith file	204
69	not a current WordSmith file	204
70	nothing activated	204
71	original text file needed but not found	204
72	printer needed	205
73	registration code in wrong format	205
74	registration is not correct	205
75	short of memory	205
76	source folder file(s) not found	205
77	stop list file not found	205
78	stop list file not read	205
79	tag file not found	206
80	tag file not read	206
81	this function is not yet ready	206
82	this is a demo version	206
83	this program needs Windows 98 or greater	206

84	to stop getting this message ...	206
85	too many requests to ignore matching clumps	206
86	too many sentences	206
87	truncating at xx words -- tag list file has more	206
88	two files needed	207
89	unable to merge Keywords Databases	207
90	why did my search fail?	207
91	word list file is faulty	207
92	word list file not found	207
93	WordList comparison file is faulty	207
94	WordSmith Tools already running	207
95	WordSmith Tools expired	207
96	WordSmith version mis-match	208
97	XX days left	208
	Index	209

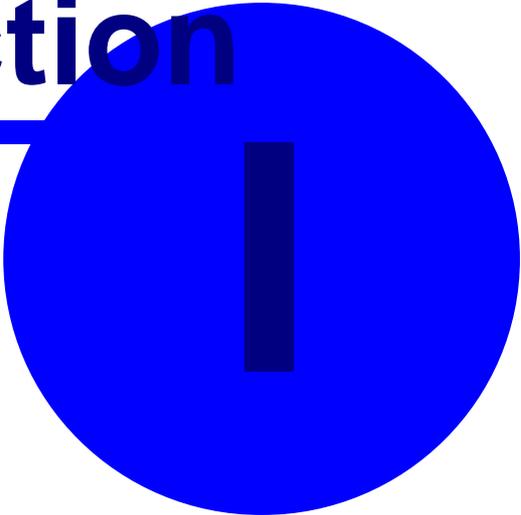
Foreword

This is just another title page
placed between table of contents
and topics

WordSmith Tools

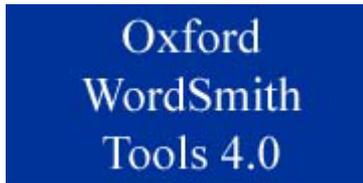
WordSmith Tools

Section



I

1 WordSmith Tools



Oxford WordSmith Tools is an integrated suite of programs for looking at how words behave in texts. You will be able to use 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 are 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

[If you have an Internet connection, click here for step-by-step screenshots showing what WordSmith does.](#)

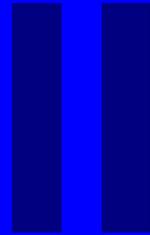
Most of the menus and dialogue boxes have help options. You can often get help just by pressing F1 or , or by choosing Help (at the right hand side of most menus). Within a help file (like this one) you may find it easiest to click the Search button and examine the index offered, or else just browse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

Click here to get started straight away with [WordList](#), [Concord](#), or [KeyWords](#).

WordSmith Tools

Overview

Section



2 Overview

2.1 What's new in version 4

Version 4 is a complete new re-write.

New features:

- [online set-by-step screenshots showing how to make wordlists, concordances etc.](#)
- [full online version of this help file](#)
- [virtually unlimited](#) concordances and word lists
- improved word list [cluster](#) handling
- variable size Concord [clusters](#)
- Concord shows where in the [sentence, paragraph, heading, and section](#) each occurrence comes
- [Unicode](#) handling of text allowing many more [languages](#)
- make a concordance and play the corresponding [sound file](#)
- enhanced [tag handling](#)
- build up your own [text corpus from the Internet](#)
- enhanced statistical functions for [collocation](#)
- advanced lemmatisation through use of [own .dll files](#)
- use of [.zip files](#)
- more Utility tools
- [version](#) checking
- more languages in [Aligner](#)

2.2 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](#).

For a step-by-step view with screenshots, click [here to visit the WordSmith website](#).

2.3 Concord



Concord is a program which makes a [concordance](#) using [DOS](#), [Text Only](#), [ASCII](#) or [ANSI](#) 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](#) for later use, edited, printed, copied to your word-processor, or saved as text files.

See also: Concord Help Contents Page, [The buttons](#)

2.4 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](#).

Listings can be [saved](#) 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](#) program.

See also: KeyWords Help Contents Page, [The buttons](#)

2.5 WordList



This program generates word lists based on one or more [ANSI](#) or [ASCII](#) text files. Word lists are shown both in alphabetical and frequency order. They can be [saved](#) for later use, edited, printed, copied to your word-processor, or saved as text files.

See also: WordList Help Contents Page, [The buttons](#)

2.6 Utilities

2.6.1 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 Oxford WordSmith Tools.

See also: [Choose Language Tool](#)

2.6.2 File Utilities



Programs to

- [compare two files](#)
- [cut large files into chunks](#)
- [find duplicate files](#)
- [rename](#) multiple files
- find "[holes](#)" in text files
- split large files into their component texts
- [join up](#) a lot of small text files into merged text files

2.6.3 Minimal Pairs



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

See also : [aim](#), [requirements](#), [choosing your files](#), [output](#), [rules and settings](#), [running the program](#).

2.6.4 Splitter



Splitter is a utility 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, [The buttons](#)

2.6.5 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.

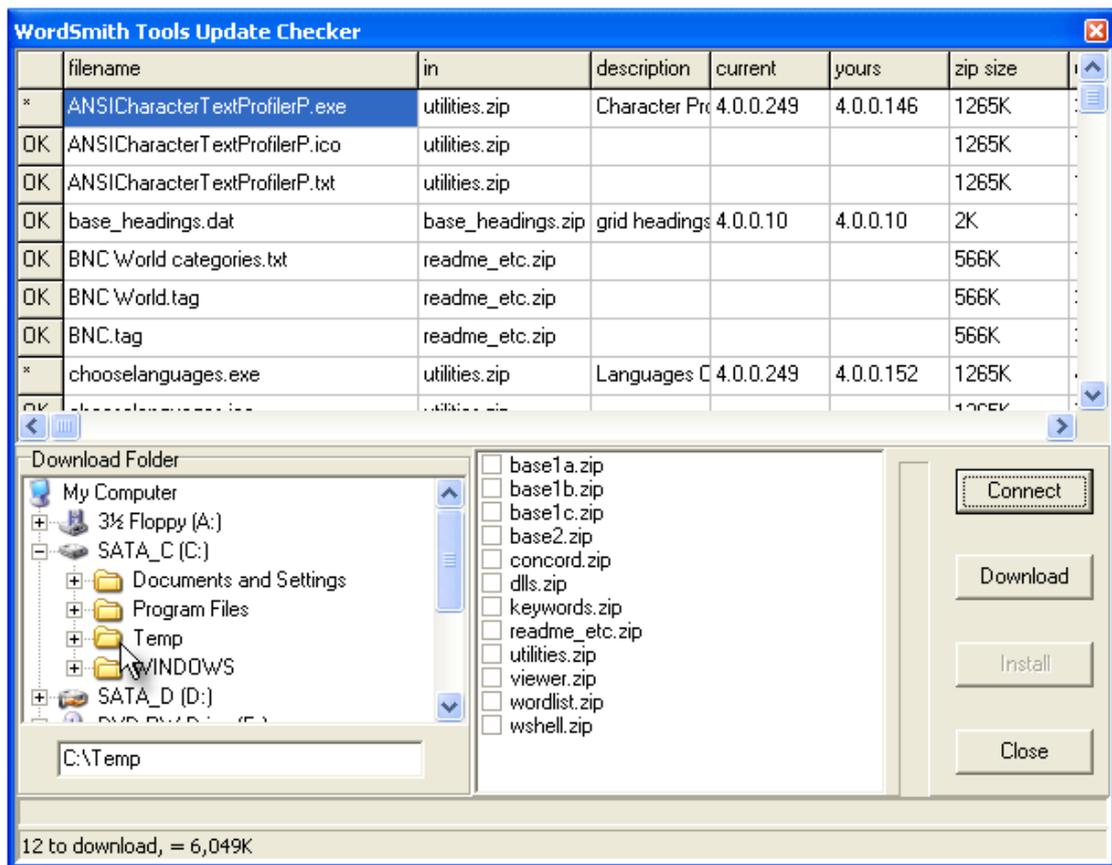
See also: Text Converter Help Contents Page, [Menu & button options](#)

2.6.6 Version Checker

4.0

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, after pressing *Connect*, you see something like this:



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 but you're free to choose somewhere else as in the screenshot where `c:\temp` has been chosen. Press *Download* if you wish to get the updated files.

	filename	in	description	current	yours	zip size
OK	Text_Converter.exe	utilities.zip	Text Converter	4.0.0.222	4.0.0.147	1246K
OK	viewer.exe	viewer.zip	Viewer and A	4.0.0.222	4.0.0.150	207K
OK	Teacher.exe	utilities.zip	Instructions fo	4.0.0.222	4.0.0.146	1246K
OK	updater.exe	utilities.zip	Registration b	4.0.0.222	4.0.0.147	1246K
OK	MinimalPairs.exe	utilities.zip	Minimal Pairs	4.0.0.222	4.0.0.146	1246K
OK	keywords.exe	keywords.zip	KeyWords too	4.0.0.222	4.0.0.148	266K
OK	File_Uilities.exe	utilities.zip	File Utilities to	4.0.0.222	4.0.0.148	1246K
OK	concord.exe	concord.zip	Concord tool	4.0.0.222	4.0.0.183	287K
OK	convertWS1to3Data.exe	utilities.zip	Converter for	4.0.0.222	4.0.0.146	1246K
OK	ANSICharacterTextProfilerP.exe	utilities.zip	Character Pro	4.0.0.222	4.0.0.146	1246K
OK	chooselanguages.exe	utilities.zip	Languages C	4.0.0.222	4.0.0.152	1246K

Download Folder

My Computer

- 3½ Floppy (A:)
- SATA_C (C:)
 - Documents and Settings
 - Program Files
 - Temp
 - WINDOWS
- SATA_D (D:)
- DVD-RW Drive (E:)
- IDE_F (F:)

C:\Temp

- base2.zip
- wordlist.zip
- wshell.zip
- readme_etc.zip
- dlls.zip
- utilities.zip
- viewer.zip
- keywords.zip
- concord.zip
- base1c.zip
- base1b.zip
- base_headings.zip
- base1a.zip

Connect

Download

Install

Close

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.

2.6.7 Viewer



Viewer 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 can also be used simply to produce a copy of a text file with [numbered sentences or paragraphs](#) or for [aligning](#) two versions of a text, showing alternate paragraphs or sentences of each.

See also: Viewer Help Contents Page, [The buttons](#)

2.6.8 Webgetter



A tool to gather text from the Internet.

The point of it...

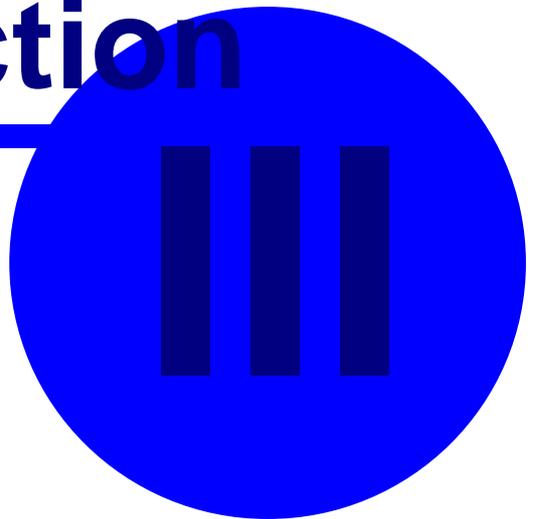
The idea is to build up your own corpus of texts, by downloading web pages with the help of a search engine.

See also: [A fuller overview](#), [Settings](#), [Display](#), [Limitations](#)

WordSmith Tools

Getting Started

Section



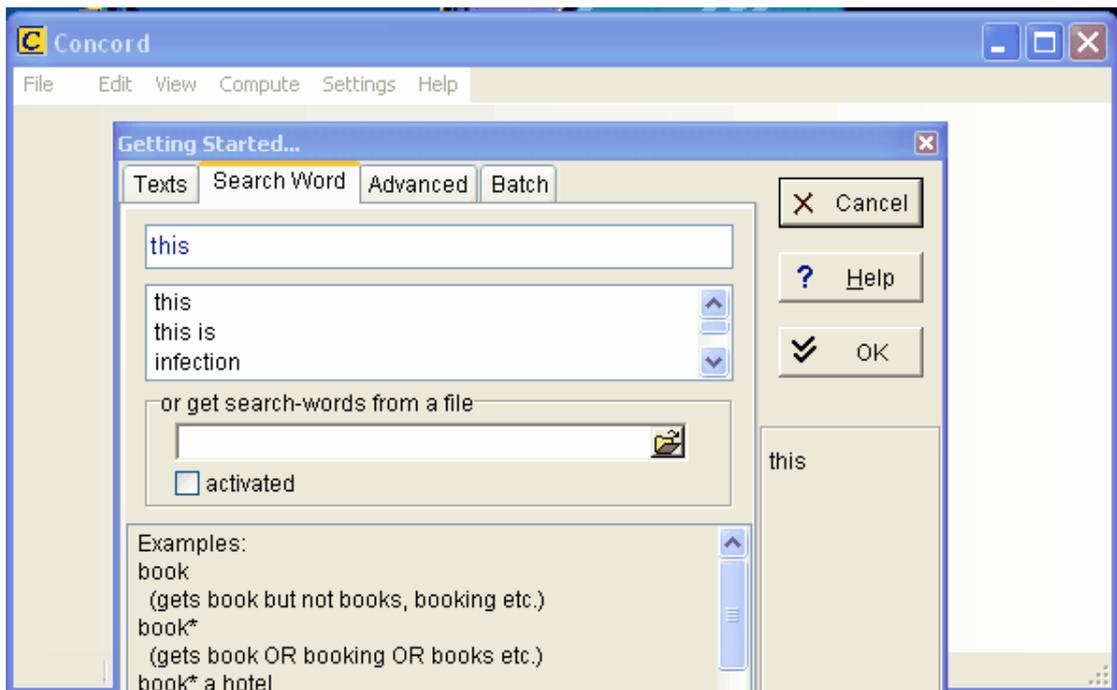
3 Getting Started

3.1 getting started with Concord

For a step-by-step view with screenshots click [here to visit the WordSmith website](#).

In the main Oxford WordSmith Tools window (the one with Oxford WordSmith Tools [Controller](#) 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.

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



(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](#) 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](#) and then press OK ().

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



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](#), 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](#).

Don't forget to [save the results](#) (press F2 or ) if you want to keep the concordance for another time.

See also: Concord Help Contents.

3.2 getting started with KeyWords

For a step-by-step view with screenshots click [here to visit the WordSmith website](#).

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

Start

You see a dialogue box which lets you [choose your wordlists](#). You'll need to choose two word lists to make a key words list from: one based on a single text, and another one based on several 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](#) 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](#) 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.

Don't forget to [save the results](#) (press F2) if you want to keep the keyword list for another time.

See also: KeyWords Help Contents, [What's it for?](#)

3.3 getting started with WordList

For a step-by-step view with screenshots click [here to visit the WordSmith website](#).

I suggest you start by trying the Wordlist program. In the main Oxford WordSmith Tools window (the one with Oxford WordSmith Tools [Controller](#) in its title bar), choose the Tools option, and once that's opened up, you'll see Wordlist. Click and WordList will open up, on the right hand side of your screen.

Start

You will see a dialogue box which lets you [choose your texts](#) or change your choice, and make a new word list.

If you have [never used WordSmith](#) 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](#).

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

Don't forget to [save the results](#) (press F2 or ) if you want to keep the word list for another time.

See also: WordList Help Contents.

WordSmith Tools

Installation and Updating

Section

IV

4 Installation and Updating

4.1 installing WordSmith Tools

You will need just under 20 Mb of space on your hard disk for the programs, but during installation you will need double that. (Anyway Windows won't run well without at *least* 20Mb spare.)

1. You have received or downloaded one or more **.exe** files. Put them in a temporary folder, e.g. `c:\temp`. It's easiest if this is a clean folder without any other files in it.
2. Run them. This will expand all the files needed for **Oxford WordSmith Tools** into the folder of your choice (`c:\wsmith4` by default).
3. Now run `c:\wsmith4\wshell.exe` to get started. You will be asked to "update from demo" and if you have the registration key you can type it in; otherwise **WordSmith** will go through its paces as a [Demonstration Version](#).
4. If you're short of disk space, you can delete the **.exe** files in `c:\temp`.

You can get a more recent version at the [WordSmith home page](#).

To un-install, just delete all the files in your `\wsmith4` folder.

See also: [Setting default options](#), [Contact Addresses](#).

4.2 network defaults

If you have bought a site licence, it's much easier to install one copy of WordSmith on a server which is accessible by all your users. Naturally, you won't want them to save any results or alter the original copy of WordSmith in that main location. So, take a look at `wshell.ini`: 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:\wsmith4
```

(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 network-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 a network installer will be a lot easier if the drive and folder you specify is truly one your users can write to!

See also: [Class Instructions](#)

4.3 version checking

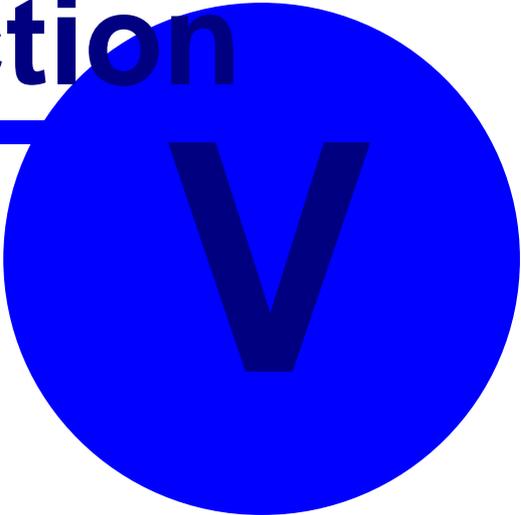
WordSmith comes with a file called `wordsmith_version_check.exe` which enables you to check whether your version is current and if not to download the necessary upgrades and patches. In order to install these, WordSmith itself will need to close down.

See also: [version information](#), [version updating](#).

WordSmith Tools

Controller

Section



V

5 Controller

5.1 adjust settings

The main Adjust Settings window in the [Controller](#). To get there, choose *Settings | Adjust Settings...* in the main Controller window.

Enables you to choose and [save](#) settings concerning:

- [font](#)
- [colours](#)
- [folders](#)
- [tags](#)
- [general settings](#)
- [match-lists](#)
- [stop lists](#)
- [lemma lists](#)
- [text and language settings](#)
- [Concord Settings](#)
- [KeyWords settings](#)
- [WordList settings](#)
- [advanced user specific settings](#)
- [index file settings](#)

5.2 accents

This window shows the accented characters available for your currently-selected [language](#).

5.3 add notes

This allows you to jot down some notes to [save](#) with your data.

For example, if you have done a concordance and sorted it carefully using your own [user-defined categories](#), you will probably want to list these and save the information for later use.

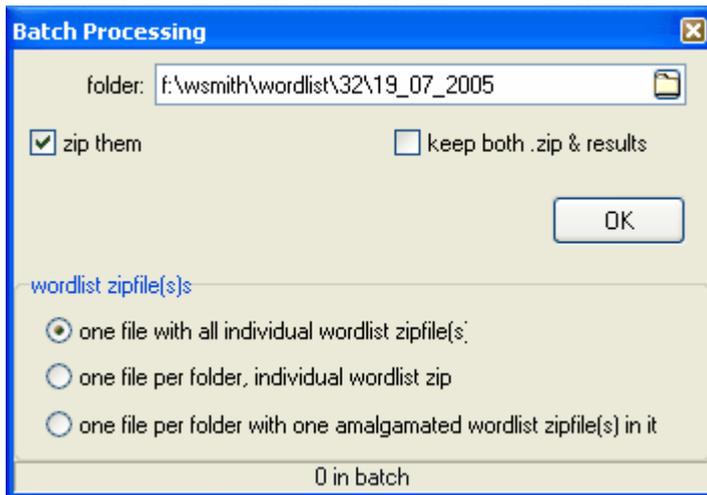
If you need access to these notes outside **Oxford WordSmith Tools**, select the text using Shift and the cursor arrows or the mouse, then copy it to the [clipboard](#) using *Ctrl-Ins* and paste into a word processor such as **notepad**.

5.4 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.



Folder where they end up

The name suggested is today's [date](#). 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 wordlists in it. Each wordlist 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`

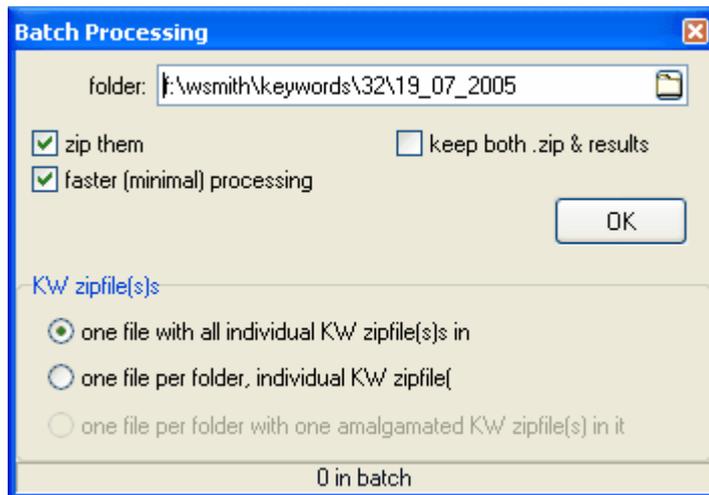
`\..\BNC\spoken`

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 wordlist or concordance from that whole folderful of texts.

Faster (Minimal) Processing



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](#), [patterns](#) and [dispersion plots](#) will be computed when each concordance is done. In KeyWords, likewise, there will be [dispersion plots](#), [link](#) 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](#) 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.

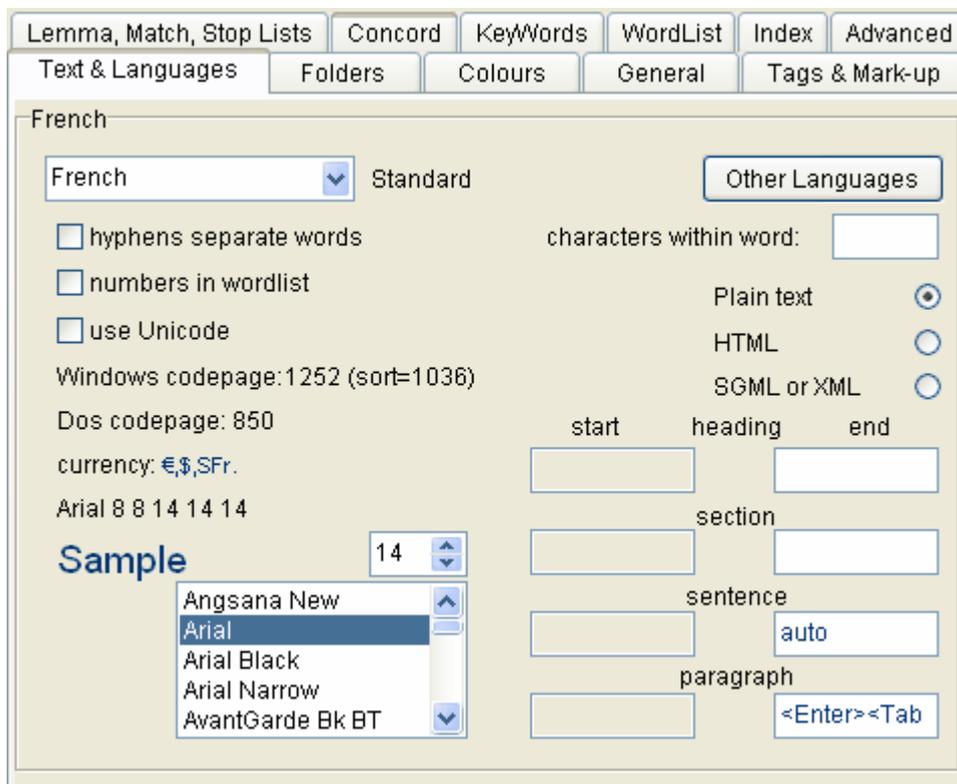
Note: of course Concord will only succeed in opening a concordance and KeyWords a key word list file. If you choose a .zip file which contains something else, it will give an error message.

5.5 batch folders

The operating system gets unhappy if there are too many files in a folder. So WordSmith batch processing creates a numbered sub-folder ...\\0 which receives the first 500 or so files; if you have chosen to work with more, it then makes another folder ...\\1 for another 500 or so, until all your lists have been made.

5.6 change font

Found in main Settings menu in all Tools or via *Adjust Settings | Text & Languages* in the [Controller](#). Enables you to choose a preferred Windows font and point size for the display windows and [printing](#) in all the Oxford WordSmith Tools suite. Note that each [language](#) can have its own different default font.



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](#) option .

Oxford WordSmith Tools will offer fonts to suit the [language](#) chosen in the top left box. Each language can have its own default font. Language choice settings once saved can be seen (and altered, with care) in `\wsmith4\language_choices.ini`.

5.7 choose favourite texts

save favourites

Used to save your current selection of texts. Useful if it's complex, e.g. involving several different folders.

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

get favourites

Used to read a previously-saved selection from disk.

By default the [filename](#) 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.

See also: [Choosing Texts](#)

5.8 choose language

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

Choose the language for the text you're analysing in the [Controller](#) under Adjust Settings | Text & Languages. The language and [character set](#) must be compatible, e.g. English is compatible with Windows Western (1252), DOS Multilingual (850).

Oxford WordSmith Tools handles a good range of languages, ranging from Albanian to Ukrainian. Chinese, Japanese, Arabic etc. are handled in [Unicode](#). You can view wordlists, concordances, etc. in different languages at the same time.

The point of it...

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 wordlists 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 wordlists (e.g. in KeyWords) relies on sort order to get satisfactory results -- you will get strange results in this if you are comparing 2 wordlists which have been declared to be in different languages.

How Languages are chosen

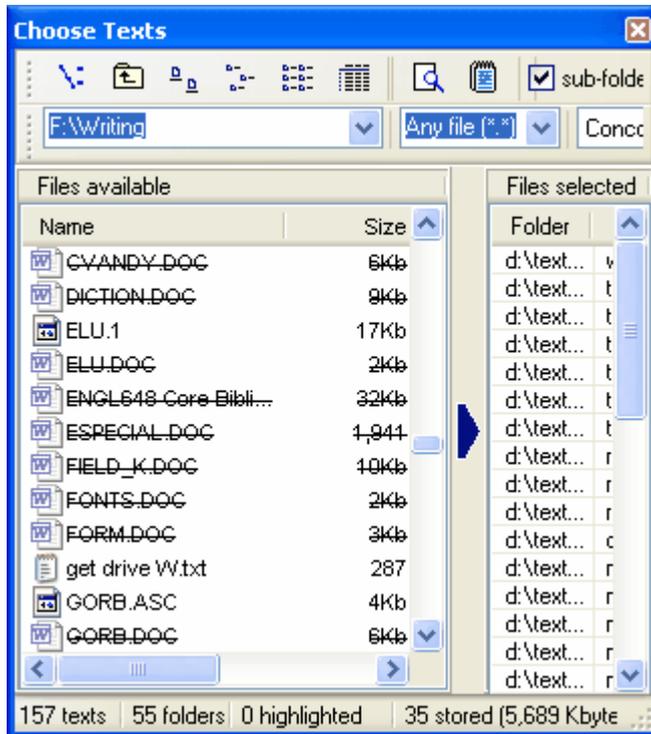
Run the Languages Chooser Utility in the main Controller Tools menu.

See also: [Choosing Accents & Symbols](#), [Accented characters](#), [Processing text in Chinese](#) etc.

5.9 choose texts

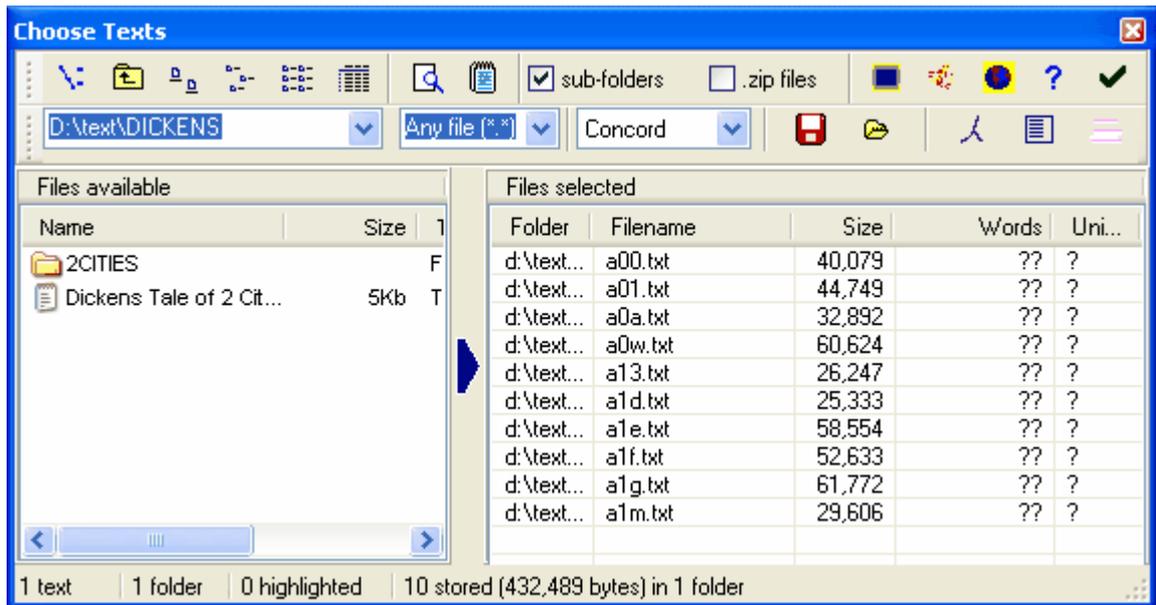
Type of text files

In WordSmith you need [plain text files](#), such as you get if you save a Word .doc as Plain Text (.txt). Any Word .doc files will look crossed out.



How to get here

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



The two main areas at left and right are

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

The big blue arrow is where 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](#).

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

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](#).

File Types

The default file specification is *.* (i.e. any file) but this can be altered in the window above the big blue arrow or set permanently in [wshell.ini](#).

Tool

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

Sorting

By clicking on *Name*, *Size*, *Type*, *Words*, *Unicode* or *Modified* you can re-sort the listing. The red and yellow button (🎲) re-orders the files (on both sides) in random order.

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.

View

Allows you to browse within the currently selected file so as to check whether to include it. Any accented characters (e.g. æ, é) or currency symbols such as £, ¥, ¢, and [tags](#) will appear according to current [Text Characteristics settings](#). You can change these while [viewing](#) the file.

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](#) so as to download text from the Internet.

Zip files

If you double-click on a [zip file](#) you can enter that as if it were a folder and see the contents. You can view these too.

Favourites

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

Test for Unicode

This button tests 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 plain ASCII or [Ansi](#) text, if it's a Word .doc file, D.

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.

OK

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*.

See also : [Finding source texts](#).

5.10 choosing files from standard dialogue box

The dialogue box here is very similar to the one used for [choosing text files](#); it also allows you to choose from a [zip file](#).

You can use [Viewer](#) 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*

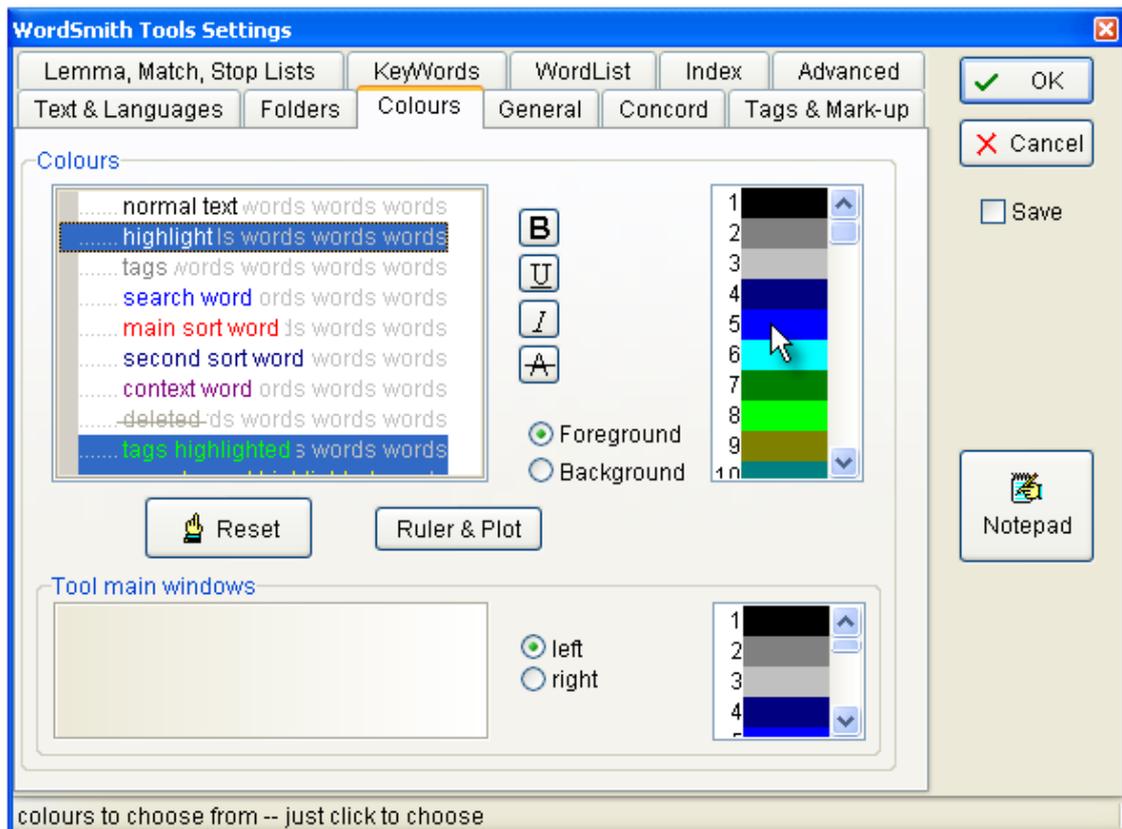
in the same zip file as your concordance or word lists.

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.

5.11 colours

Found in main Settings menu in all Tools and Adjust Settings in the [Controller](#). 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	mark-up
search word	concordance search word; words in (key) word lists
main sort word	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	concordance search word when selected
highlighted	
main sort word	first sort when selected
highlighted	
second sort word	first tie-breaker sort when selected
highlighted	
context word	context word when selected
highlighted	
most frequent	most frequent collocate or detailed consistency word, p value
collocate	in keywords
viewing texts	in the text viewer

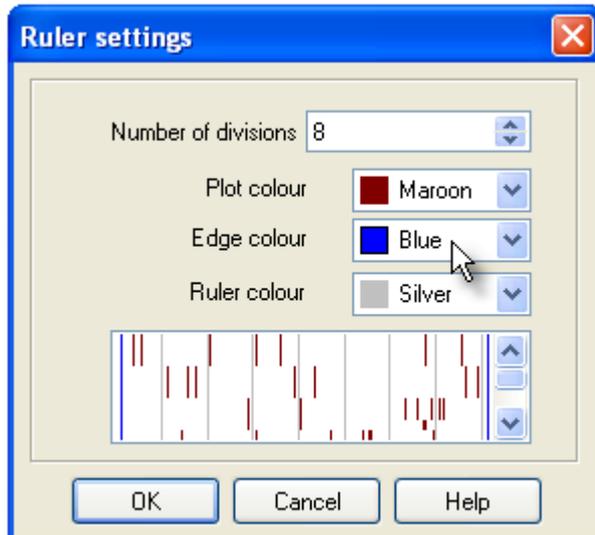


To alter colours, first click on the wording you wish to change (you'll see a difference in the left margin: here highlight has been chosen), then click on a colour in the colour box (where the cursor is in the screenshot). The radio buttons below the colours determine whether you're changing foreground 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](#) 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).

Ruler and Plot

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



See also: [layout](#) for changing the individual colours of each column of data.

5.12 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	0.09	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).

5.13 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.

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

Rel(2) ÷ 5 for each row in your data, the new column will contain the data from column 2 of the same row, divide it by 5, and put the result in your new column.

RelC(2) for each row in your data, the new column will contain the data from column 2 of the same row, add it to a running total, and put the result in your new column.

Rel(3) + (Rel(2) ÷ 5) for each row in your data, the new column will contain the data from column 2 of the same row, divide it by 5, add it to the data from column 3 of the same row, and put the result in your new column.

Abs(2;1) ÷ 5 for each row in your data, the new column will contain the data from column 2 of row 1, divide it by 5, and put the result in your new column. 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 for each row in your data, the new column will contain the data from column 2 of the same row, divide it by column 2 of row 1 and multiply it by 100, putting the result in your new column. 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](#).

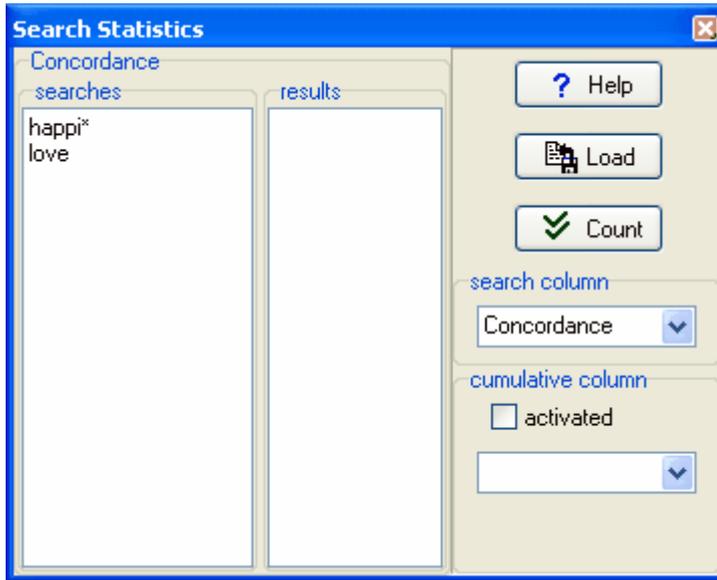
See also: [count data frequencies](#), column totals

5.14 count data frequencies

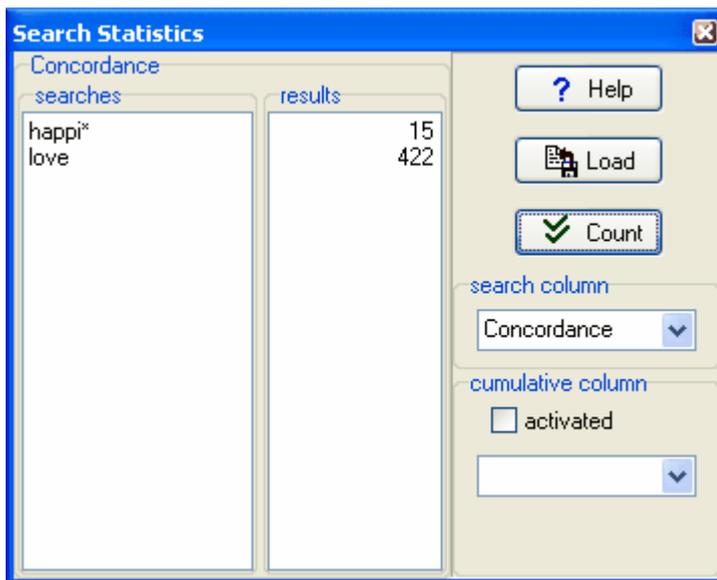
You may want to know how many of your concordance lines contain "**happi***" or how many items in a word-list end in ***1y**. To do this, choose *Summary Statistics* in the *Compute* menu.

An Example

You have a concordance already computed. Select anywhere in the concordance lines and choose *Compute : Summary Statistics*. Type **happi*** and **love** in the searches window.



Press Count -- you should see something like this:



The procedure has processed all your concordance lines and found out that 15 contain **happi*** and 422 contain the whole word **love** (not **loved**, **loves**).

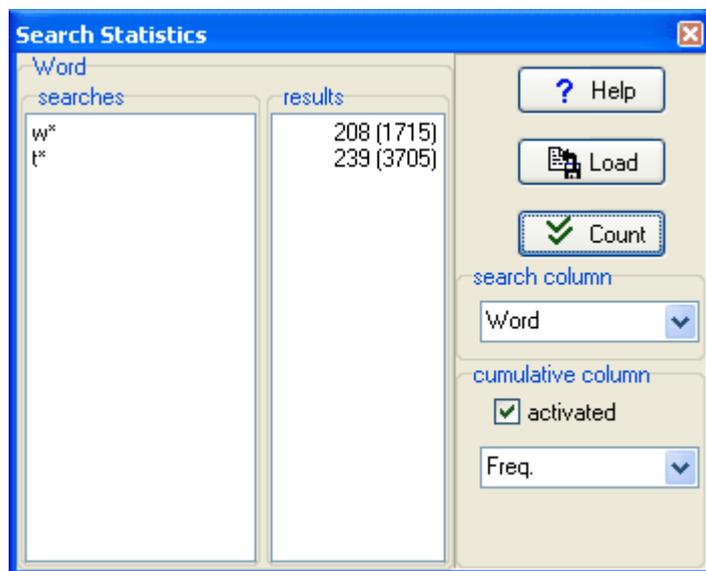
Search Column

This combobox lets you choose which column of data to count in.

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 combobox are of the numerical data only. Select one and ensure *activated* is ticked.

In this example, a word-list was computed and a search was made of words beginning with **w** or **t**. There are 208 of those beginning with **w**. In brackets you can see 1715 -- this means that the cumulative count of words beginning with **w** in terms of their frequency (*Freq.* column) is 1715, in other words an average frequency of about 8 (1715 / 208). But for the words beginning in **t**, although the absolute number is similar (239), the average cumulative frequency is about 15. That is because there are lots of high-frequency words beginning in **t** in English.



Load

This allows you to load into the searches window any [plain text](#) file which you have prepared previously.

See also: [compute new column of data](#).

5.15 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+Ins. This puts it into the clipboard.

If you press Copy (this needs updating) you get various choices:

[clipboard](#)

[saving as a text file](#)

[printing](#)

[save](#) as data (not the same as saving as text: this is saving so you can access your data again another day)

all: copy all the rows and columns of data

selected: copy only the rows and columns which you've already highlighted

specify: copy according to criteria which depend on the Tool:

In the case of **Concord**, you can [save](#) concordance lines which you have classified [according to your own categories](#). You can specify one category (between **a** and **z** or between **A** and **Z**) to save.

5.16 custom processing

This feature -- which is not for those without a tame programmer to help -- is found under Adjust 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;
```

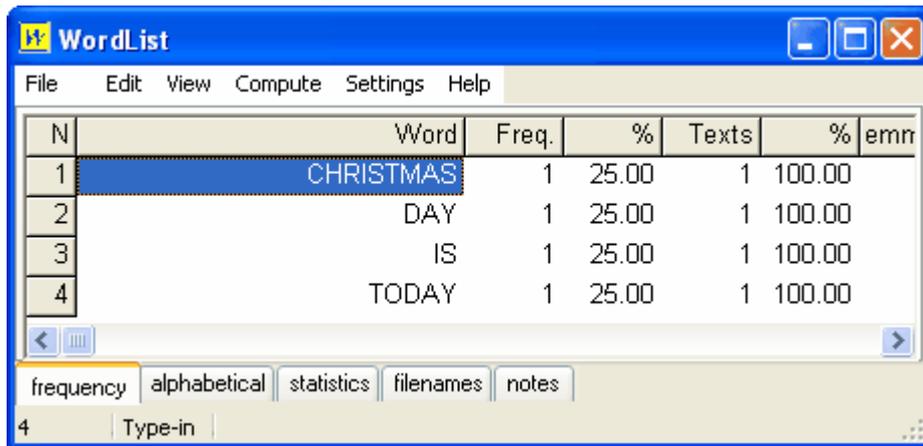
The language_identifier is a number corresponding to the language you're working with. See [List of Locale ID \(LCID\)](#)

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.



The source code for the .dll in Delphi is this

```
*****
```

```
library WordSmith4CustomDLL;
```

```
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,
      PChar(original), -1,
      PChar('EASTER'), -1) - 2 = 0
    then
      Result := pchar('CHRISTMAS');
  end;
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;

function HandleConcordanceLine
(source_line : pointer;
 hit_position,
 hit_length : word;
 byte_position_in_file,
 language_id : DWORD;
 is_Unicode : WordBool;
 filename : pchar) : pointer; stdcall;

function extrasA : string;
begin
  Result := #9+pchar(filename)+
           #9+ IntToStr(byte_position_in_file)+
           #9+ IntToStr(hit_position)+
           #9+ IntToStr(hit_length);
end;

function extrasW : widestring;
begin
  Result := extrasA;
end;

var f : TextFile;

  output_file : string;
begin
  Result := source_line;
  output_file := ChangeFileExt(ParamStr(0),'')+
  '_user_dll_concordance_lines.txt';
  if (not IsPathDelimiter(ExpandUNCFileName(ParamStr(0)),1)) and
  (DiskFree(Ord(UpCase(output_file[1]))-64) > 1024*2000) then
  try
    if FileExists(output_file) then begin
      AssignFile(f,output_file);
      Append(f);
    end else begin
      AssignFile(f,output_file);
      Rewrite(f);
    end;
    if is_Unicode then
```

```

        Writeln(f,pwidechar(source_line)+extrasW)
    else
        Writeln(f,pchar(source_line)+extrasA);
    Flush(f);
    CloseFile(f);
except
end;
end;

exports

    ConcordChangeWord,
    KeyWordsChangeWord,
    WordlistChangeWord,
    HandleConcordanceLine;

begin
end.

```

5.17 customising popup menus

Logging

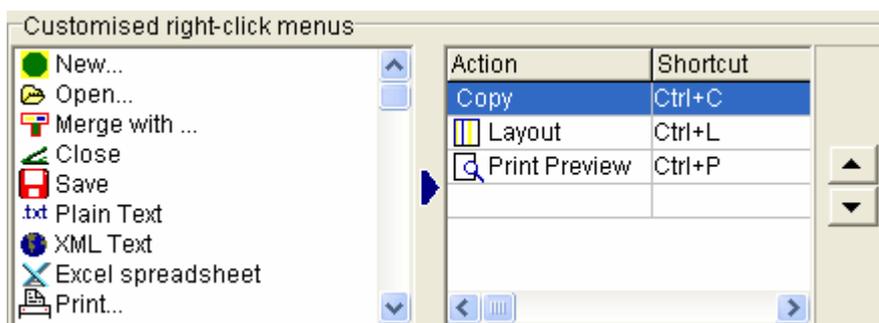
Logging is useful if you are getting strange results and wish to see details of how they were obtained. If this is enabled, WordSmith will save some idea of how your results are progressing in the log-file.

In the main WordSmith Controller program, choose *Adjust Settings | Advanced*. 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. See also: [emailed error reports](#).

Customising popup menus

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 *Adjust Settings | Advanced*.



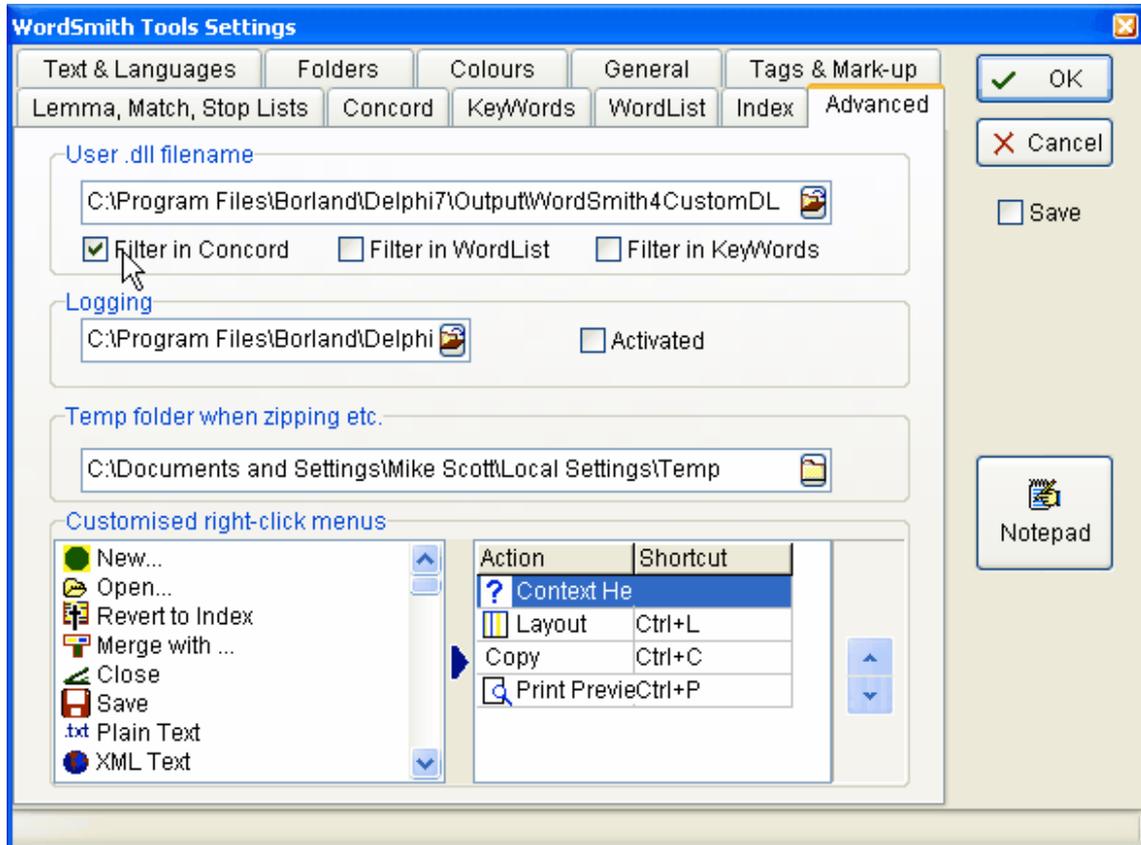
You will see a list of menu options at the left, and can add to the list on the right by selecting one on the left and pressing the button in the middle. To remove a choice in the list on the right, select it and press *Del*. To re-order the choices, press the up or down arrow.

Whatever is in your popup menu will also appear in the [Toolbar](#).

To save the choices permanently, see [Saving Defaults](#).

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 `\wsmith4` folder, if there is space on the hard disk.



See also : [menu and button options.](#)

5.18 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	

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



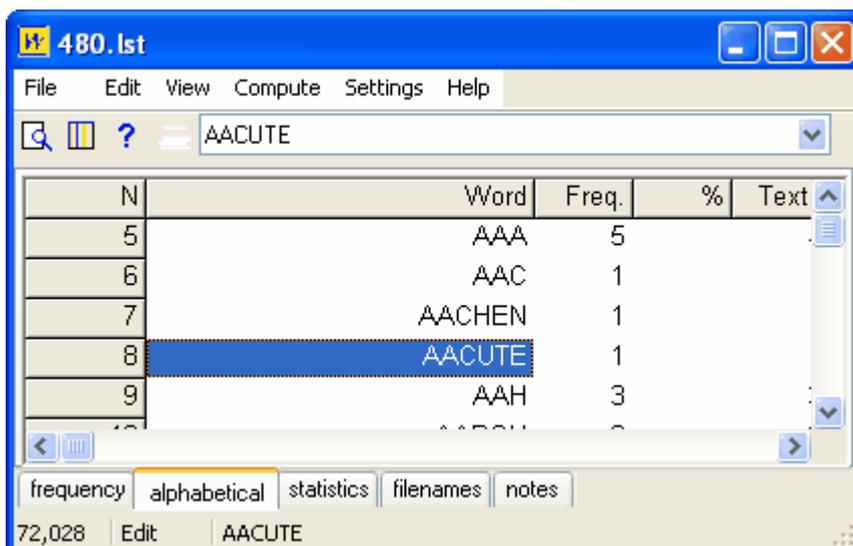
and can change the options for this list to *Set* (to classify your words, eg. as adjectives v. nouns) or *Edit*, to alter them. Note that some of the data is calculated using other data and therefore cannot be edited. For example, frequency percentage data is based on a word's frequency and the total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

Choose *Edit*.

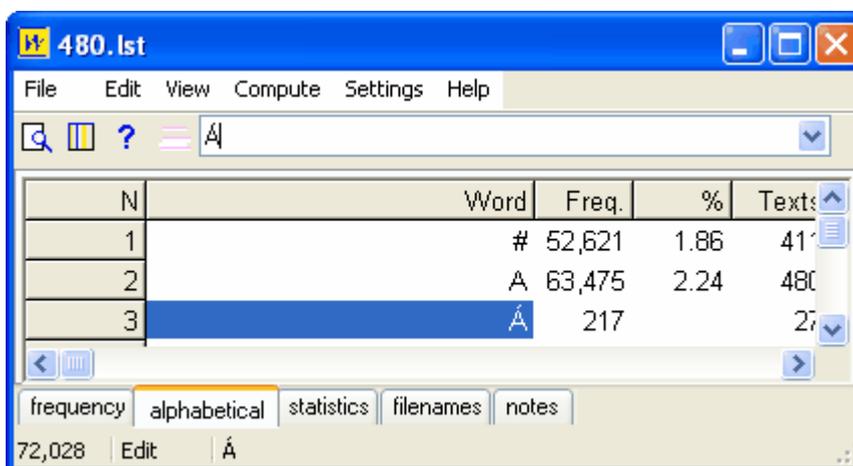
Now, in the column which you want to edit, press any letter.

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**.



If you now type \acute{A} , you will immediately see the result in the window:



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.



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

See also: [joining entries](#), [finding source files](#).

5.19 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  option, e.g. a [key words](#) listing.

It enables you to seek out all text files which contain **at least one** mention of **each** of the words you have marked (with ). Before you click, [choose the set of texts](#) which you want to peruse.

What you get

A concordance based on all the words you marked, showing which [text files](#) they were found in. But it is a "fussy" concordance: any text files which doesn't have *all* the words you selected get ignored.

5.20 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 `\wsmith4` folder where the WordSmith programs are or in the "instructions folder" explained under [Network Defaults](#). 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.

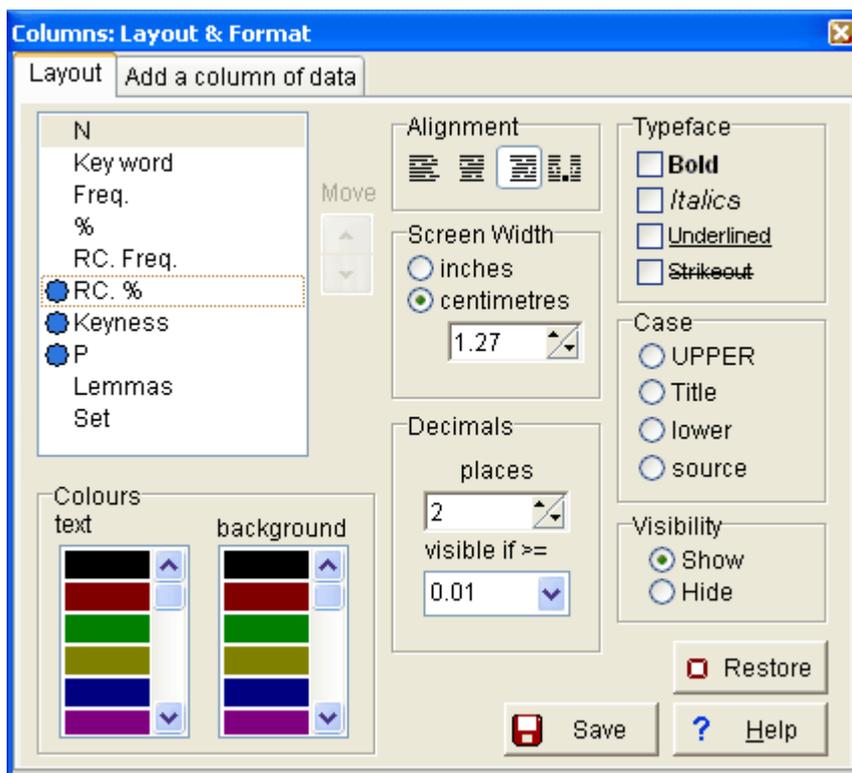
See also: [Network defaults](#)

5.21 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](#).



In the top left window of Layout, you see the available column headings. 3 of the headings have been activated (by clicking) so that settings can be changed:

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).

Case

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

Decimals

the number of decimal places for numerical data, where applicable.

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 layouts (see below).

Save

The point of this Save option is to set all future lists to a preferred layout. Suppose you have a concordance open. If you change the layout as you like and [save](#) 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 `\wsmith4` folder, eg. `Concordance list customised.dat`.

Alternatively you can choose always to show or hide certain columns of data with settings in your `wshell.ini` file. For example, in the [Concord] section of `\wsmith4\wshell.ini`, to avoid seeing the Set column, you would change

```
show set column=YES  
to  
show set column=NO
```

See also: [setting & saving defaults](#), setting [colour](#) choices in Oxford WordSmith Tools [Controller](#).

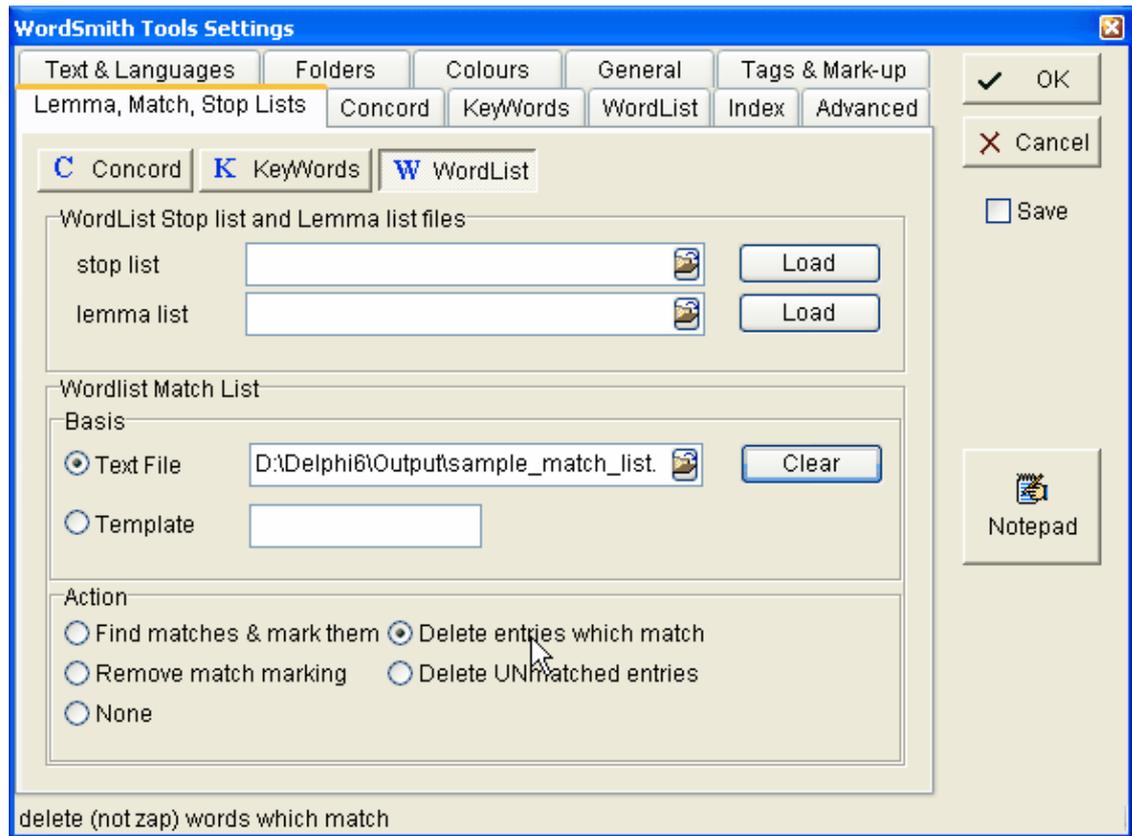
5.22 match words in list

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 (wordlist, 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

Click in the column whose data you want to match up. This will usually be one showing words, not numbers. Then press the *Match List* button. The main Controller settings dialogue box appears.



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**). If you choose a file, the Controller will then read it and inform you as to how many words there are in it.

Action

Remember to set the Action to something suitable. The cursor in the screenshot above shows that the action wanted here is to delete any matched words.

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 the match list included **is** and **THIS**).

The screenshot shows a window titled 'teststop_list.lst' with a menu bar (File, Edit, View, Compute, Settings, Help) and a table of data. The table has columns: N, Word, Freq., %, Texts, %, emn. The data rows are as follows:

N	Word	Freq.	%	Texts	%	emn
9	IN	1	5.56	1	100.00	
10	IS	1	5.56	1	100.00	
11	IT	1	5.56	1	100.00	
12	ON	1	5.56	1	100.00	
13	SO	1	5.56	1	100.00	
14	STOPLIST	1	5.56	1	100.00	
15	TEST	1	5.56	1	100.00	
16	THIS	1	5.56	1	100.00	

At the bottom of the window, there are tabs for 'frequency', 'alphabetical', 'statistics', 'filenames', and 'notes'. The 'frequency' tab is selected. Below the tabs, it shows '18 Type-in'.

You can obtain statistics of the matches, using the [Summary Statistics](#) menu option.

See also: [Comparing Word-lists](#), [Comparing Versions](#), [Stop Lists](#), [Lemma Matching](#)

5.23 print and print preview

This takes you by default to a print preview, which shows you what the current page of data looks like, and from which you can print.

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.

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. If you include <Date> this will put today's date. 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.

See also: [Printer Settings](#)

5.24 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](#)?

and

2) has no set of [favourite text](#) files been saved for easy retrieval?

5.25 previous lists

This window shows the list 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 alphabetically, choose *Resort*.

5.26 quit WordSmith

Alt-X is the hot key.

Closing Oxford WordSmith Tools [Controller](#) 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 Oxford WordSmith Tools. This feature can be turned off temporarily via a Wshell menu option or permanently in `\wsmith4\wshell.ini`.

5.27 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 allows you to specify how many entries you want to have in the list. If you reduce the data, entries will be randomly [zapped](#) 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](#), [editing a list of data](#).

5.28 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](#), e.g. if you simply want to insert your results into your word processor.

If you want to copy the data in colour, or export a plot, you should definitely use the [clipboard](#). 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 () , 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 none is 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 can only handle up to 65,000 rows. It will look something like this:

WordSmith Tools 4.0 -- 16/10/2004							
	Word	Freq.	%	Texts	%	Lemmas	Set
1							
2							
3	1	THE	46	8.88	1	100	
4	2	A	20	3.86	1	100	
5	3	OF	16	3.09	1	100	
6	4	AND	13	2.51	1	100	
7	5	SPLEEN	11	2.12	1	100	
8	6	TO	11	2.12	1	100	
9	7	BLOOD	9	1.74	1	100	
10	8	IN	8	1.54	1	100	
11	9	IS	8	1.54	1	100	
12	10	#	6	1.16	1	100	
13	11	WAS	6	1.16	1	100	
14	12	WITH	6	1.16	1	100	

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](#)). 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.

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

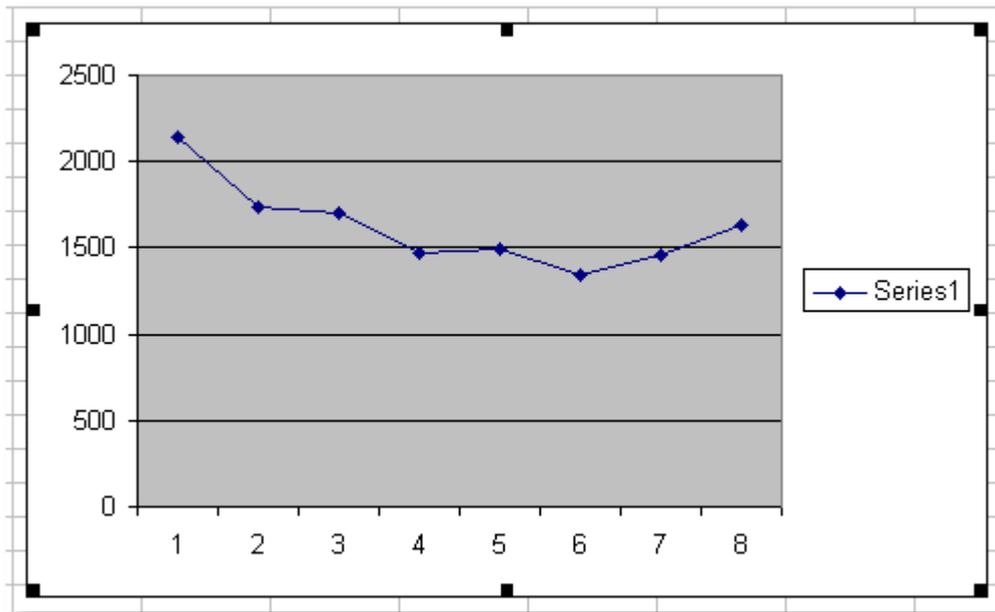
Divisions	1/8	2/8	3/8	4/8	5/8	6/8	7/8	8/8	
Raw Totals	2140	1735	1699	1466	1495	1347	1458	1632	format division data as percentage!
cep.txt	0.237705	0.155738	0.262295	0.065574			0.065574	0.213115	
hp6.txt	0.333333	0.166667		0.166667		0.333333			
bn4.txt	0.333333		0.166667	0.166667	0.333333				
asu.txt	0.142857	0.142857	0.571429	0.142857					
cbg.txt	0.146154	0.223077	0.207692				0.207692	0.215385	
g2e.txt	0.222222	0.111111	0.277778	0.111111	0.111111	0.055556	0.055556	0.055556	
cm6.txt	0.5					0.5			
a0f.txt	0.111111	0.111111		0.222222	0.111111	0.111111	0.111111	0.222222	
guy.txt	1								
k5m.txt	0.336957	0.141304	0.086957			0.086957	0.163043	0.184783	
hj4.txt	0.217949	0.166667	0.333333	0.064103			0.051282	0.166667	
h9t.txt	1								
bme.txt	0.25					0.25	0.25	0.25	
acr.txt	0.117647	0.176471	0.176471	0.058824	0.176471	0.117647	0.117647	0.058824	
cmu.txt	0.2	0.2				0.2	0.2	0.2	
hd7.txt	0.333333		0.666667						
jye.txt	0.571429	0.142857	0.142857			0.142857			
k9j.txt	0.333333	0.166667	0.333333				0.166667		
gxf.txt	0.333333	0.333333					0.333333		
b2w.txt	1								
edf.txt	1								

The plot data are divided into the number of segments set for the ruler (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.
14.67%	22.31%	20.

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



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.

5.29 save defaults

Settings can be altered by choosing *Adjust Settings* in the Oxford WordSmith Tools [Controller](#). Any setting menu item in any Tool gives you access to these:

Colours, Folders, Text, General, Tags, Stop Lists, Concord, KeyWords, WordList

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

colours	customise the default colours
folders	set WordSmith so it "knows" which folders you usually use
text	character set , treatment of hyphens & numbers, default file extension
general	restore last file , printing
tags	tags to ignore, tag file, tag file autoloading
stop lists	for Concord, KeyWords and Wordlist

matching	files to match up, or lemma files to mark lemmas in a word list, etc.
Concord	number of entries, sort system, collocation horizons
KeyWords	procedure , max. p value , database & associate minimum frequencies, reference corpus filename
WordList	word length & frequencies, type/token # , cluster settings

permanent settings and wshell.ini file

You can save your settings by checking the save box after adjusting settings. Or by editing the `wshell.ini` file, installed when you installed Oxford WordSmith Tools. This specifies all the settings which you regularly use for all the suite of programs, such as your text and results [folders](#), screen [colours](#), [fonts](#), the default [columns](#) to be shown in a concordance, etc. You can see `\wsmith4\wshell.ini` by choosing *Settings | See Current*.

show help file

In the general tab of Adjust 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 network installation.

sayings

Using Notepad, you can edit `wsmith4\sayings.txt`, which holds sayings that appear in the main [Controller](#) window, if you don't like the sayings or want to add some more.

network 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 `wshell.ini`, usually the root folder of `c:\`. The first time you use WordSmith, you will be prompted to Adjust Settings, choose appropriate [Folders](#), [Text](#) Characteristics, [Tag](#) details etc. and enable the Save checkbox, after which your settings will be saved 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 `wshell.ini`, where it may be both read and altered. If Oxford WordSmith Tools finds a copy of `wshell.ini` in that folder it will be able to use your personal preferences.

5.30 save results

To save your corrected results use *Save (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](#), first.

Saved data is in a special **Oxford 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](#) or by copying the data to the [clipboard](#).

save part of the data only

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

5.31 search & replace

Some lists, such as lists of [filenames](#), allow for searching and replacing. Like a search operation, the search operates on the *current column* of data. If you choose "Case sensitive" each letter must match in terms of upper case and lower case. You can choose to search "Down" from the next entry below the current one to the end, or else "Up" to the first entry. You can also choose to "Confirm each change" if you wish.

Whole word – or bung in an asterisk

"Whole word" means search for a word with a [word separator](#) at each end. 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**, or **book** in a two or three word entry such as **she'll book it**).

Word lists can be sorted by suffix: see [WordList sorting](#).

See also: [Searching by Typing](#), [Searching with F12](#), [Accented Characters & Symbols](#).

5.32 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](#), [Searching for a word or part of one](#), [Search & Replace](#), [WordList sorting](#)

5.33 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. If you choose "Case sensitive" each letter must match in terms of upper case and lower case. You can choose to search "Down" from the next entry below the current one to the end, or else "Up" to the first entry.

Whole word – or bung in an asterisk

"Whole word" means search for a word with a [word separator](#) at each end. 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**, or **book** in a two or three word entry such as **she'll book it**).

Word lists can be sorted by suffix: see [WordList sorting](#).

See also: [Searching by Typing](#), [Search & Replace](#), [Accented Characters & Symbols](#).

5.34 see filenames

This button enables you to open a new window, displaying the [text filename](#) 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](#), the information will be permanently recorded.

In the case of key word lists, the data comes from a word list. If the word list was based on just one text file, you'll see the text file name, but if on more than one, you'll see the name of the word list file itself: to see the original text file names, you could open up the word list and press the `filenames` button in that.

See also: [finding source files](#).

5.35 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 a file called `stoplist.stp` (in your `\wsmith4` folder) which you could use as a basis and save under a new name.

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** 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 `wshell.ini` as [defaults](#).

See [Match List](#) 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](#).

See also: [Making a Tag File](#), [Match List](#), [Lemmatisation](#).

5.36 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:

Continue

go on as if you had not interrupted anything

Finish this file, then stop

a graceful stop. 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 now

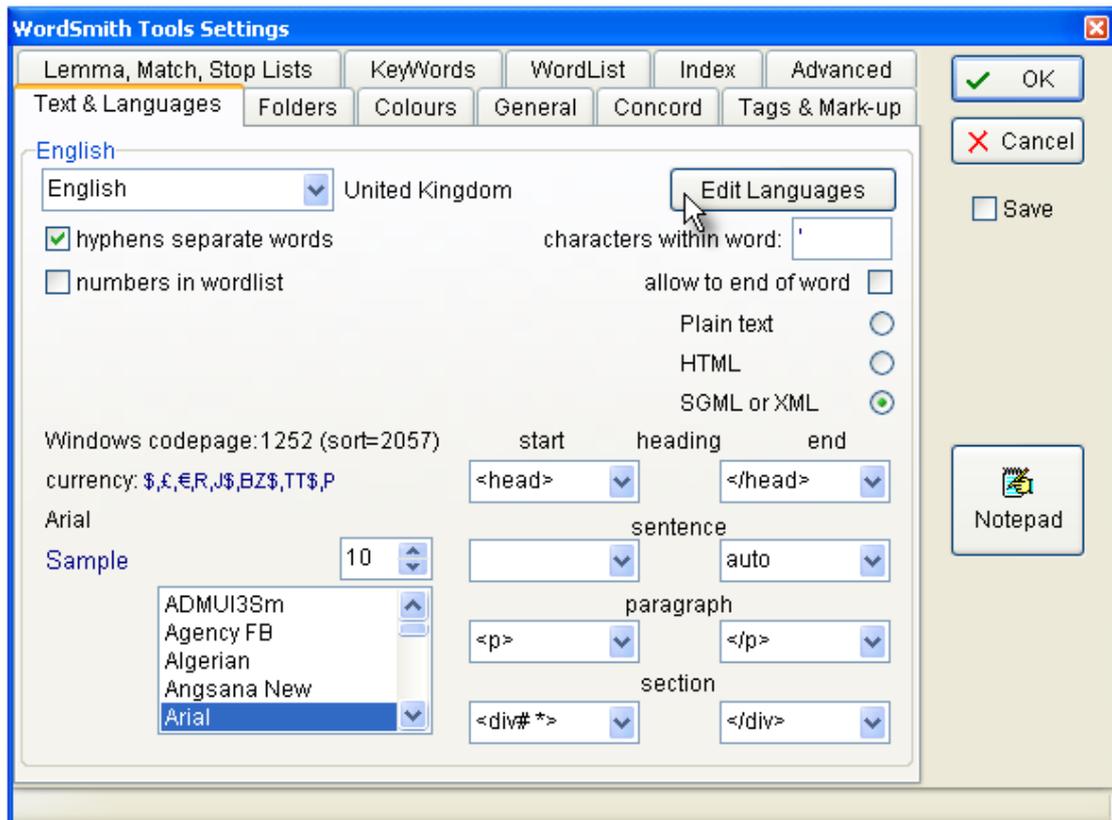
a less graceful stop, very useful if you're ploughing through massive CD-ROM files. WordSmith will stop processing the current file in the middle, but will retain any data it has got so far.

Quit this Tool

a panic stop. The whole Tool (Concord or WordList, or whatever) will close down and some system resources [memory](#) may be wasted. The [Controller](#) will not be closed down.

Press *Suspend* again to effect your choice.

5.37 text and languages



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](#) 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.

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](#) in [HTML](#), [SGML](#) or [XML](#) you should choose one of the options here. That way, the Tools can make optimum use of sentence, paragraph and heading markup.

Windows format etc.

Information about Windows [character sets](#) for the language you are working with.

start & end of heading

For the Tools to count headings, they need to know how to recognise the start and end of one. If your text is [tagged](#) 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](#) text which is not consistent, using sometimes `</h1>` and sometimes `</H1>`, then use [Text Converter](#) to make your texts consistent).

start & end of section

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

start & end of sentence

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

start & end of paragraph

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](#) 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.

which files

. means show all text types; you may prefer to limit the text files e.g. to ***.txt**.

See also: [Tagged Text](#), [Stop Lists](#), [Choosing a new language](#). [Processing text in Chinese etc.](#)

5.38 window management

The main Oxford WordSmith Tools [Controller](#) will be at the top left corner of your screen, half the screen width and half the screen height in size. With the exception of Viewer, and Concord, the main window for each Tool will appear to the right of it, and the same size. Each Tool main window will come just below any previous ones. Individual windows of results in each Tool will be restricted to that Tool's main window, and can be tiled or cascaded.

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](#) window when using individual Tools.

tile and cascade

All the main Tools show you which windows are active, listed below the item *Window* in the main menu. The current one will be ticked. To bring another one to the top, just click on the name in the list.

Or to rearrange a number of different windows, you can *Tile* them (make windows of equal sizes) if there are 5 or less, or else *Cascade* them vertically below each other. You can also *Tile* or *Cascade the Tools* from the main **Oxford WordSmith Tools** program.

screen clutter

It is easy to get a rather cluttered screen if you have several concordances, each with plot, [cluster](#), collocate and pattern windows opened up. Remember that all these windows depend on their "parent" window, the concordance itself. Likewise, a keywords plot is a "child" of a keywords listing. You can close any of them down at any time, and call them back up as long as the parent window is still open.

If you have a concordance with collocates and patterns open too, I suggest you minimise the concordance window then *Tile*; this will show the collocates and the patterns while keeping the concordance unobtrusively out of the way.

restore last file

A convenience feature: the last file you saved or retrieved will by default be restored when you re-enter Oxford 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 `wshell.ini` (in your `\wsmith4` folder).

5.39 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 (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.

See also : [reduce data to N entries](#).

WordSmith Tools

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."). **Oxford WordSmith Tools** has been designed so that you can choose what to ignore and what to see.

You may want to *translate* [HTML or SGML](#) tags or entity references: if your text has **É**; 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 **Oxford WordSmith Tools** 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](#) button. Any retained tags will be visible, and ignored tags replaced by spaces.

See also: [Handling Tags](#), [Making a Tag File](#), [Showing Nearest Tags in Concord](#), [Concord Sound and Video](#), [Tag Concordancing](#), [Types of Tag](#), [Viewing the Tags](#), [Using Tags as Text Selectors](#), [Tags in WordList](#)

6.2 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](#), [sections](#), [sentences](#) and [paragraphs](#). Type these in the appropriate spaces when selecting [Text Characteristics](#).

tags within 2 separators

These tags are often used to signal the part of speech of each word; they're also widely used in [HTML](#), [XML](#), [SGML](#) 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](#) markup):
`<Noun>`, `<Verb>`, `<Pronoun>`.

entity references

[HTML](#), [XML](#) and [SGML](#) use so-called entity references for symbols which are outside the standard alphabet, e.g. `é`; `té` which represents `té`.

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](#) 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](#) 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 Translated](#) option allows you to specify entity references which you want to convert on the fly, such as `é`.

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 ... 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](#).

See also: [Overview of Tags](#), [Handling Tags](#), [Making a Tag File](#), [Showing Nearest Tags in Concord](#), [Tag Concordancing](#), [Viewing the Tags](#), [Using Tags as Text Selectors](#), [Concord Sound and Video](#)

6.3 handling tags

ignore all tags

Specify all the opening and closing symbols in Adjust Settings | Tags | Tags 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 : `<*>` This will cut out all wording starting at each `<` symbol and ending at the next `>` symbol (up to 1,000 characters apart).

ignore some tags and retain others

If you want to ignore some but retain others, you will need to prepare a [tag file](#) which lists all those you want to keep. These will then appear in your word lists and concordances. You get Oxford 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 [filename](#) 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.

If you declare the filename in your [defaults](#) (`wshell.ini`) and include `autoload tagfile=YES`, the tags will be automatically loaded as WordSmith starts up.

translate entity references into other characters

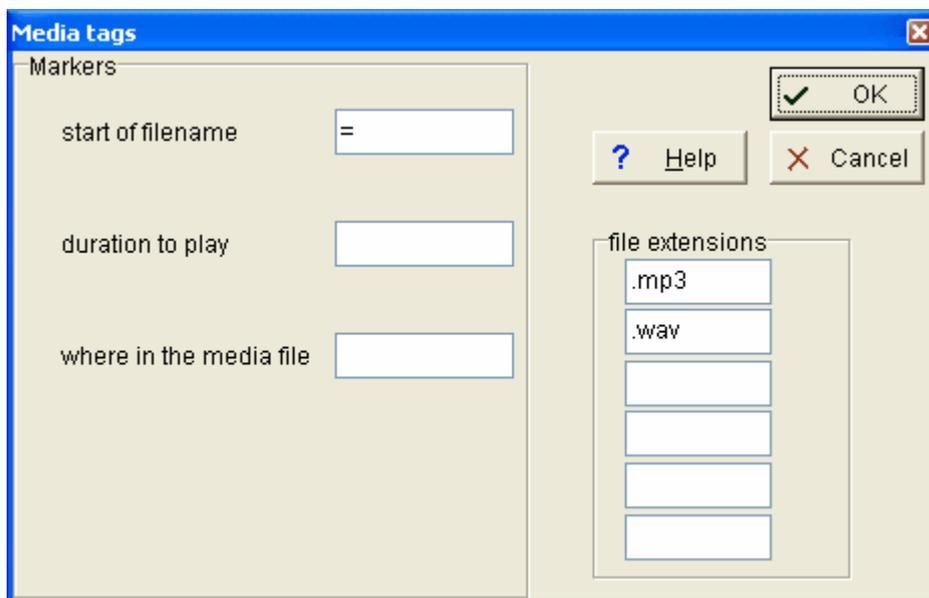
If you use [SGML or HTML](#) tagged text, you may want to translate symbols. For example, SGML, XML, HTML use `—` instead of a long dash. To do this, first prepare a [Tag File](#) which contains the strings you want to translate. Then choose *Adjust Settings | Tags & Markup | Entity File (entities to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

If you declare the filename in your [defaults](#) (`wshell.ini`) and include `autoload tags to translate file=YES`, the entities will be automatically loaded as WordSmith starts up.

See also: [Overview of Tags](#), [Making a Tag File](#), [Showing Nearest Tags in Concord](#), [Tag Concordancing](#), [Types of Tag](#), [Viewing the Tags](#), [Using Tags as Text Selectors](#), [Saving defaults](#)

6.4 multimedia tags

In this screenshot you see an example of how to define your multimedia tags. To get a reference like `` 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. The file extensions 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.

Duration and Length to Play

Duration to play and where to start playing are measured in seconds.

If they are left blank, the whole sound or video file will be played.

Your text files might have duration and start position markers, eg.

`<sound$$C:\mysounds\talk.wav start:150 play:5>`. If so, you'd specify duration to play as **play:** and where in the media file as **start:**

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".

defaults

The defaults are: play .mp3 and .wav files. Once you've completed this, [save your defaults](#) for next time.

See also: [Sound and Video in Concord](#), [Overview of Tags](#), [Making a Tag File](#), [Tag Handling](#), [Tag Concordancing](#), [Showing Nearest Tags in Concord](#), [Viewing the Tags](#), [Types of Tag](#)

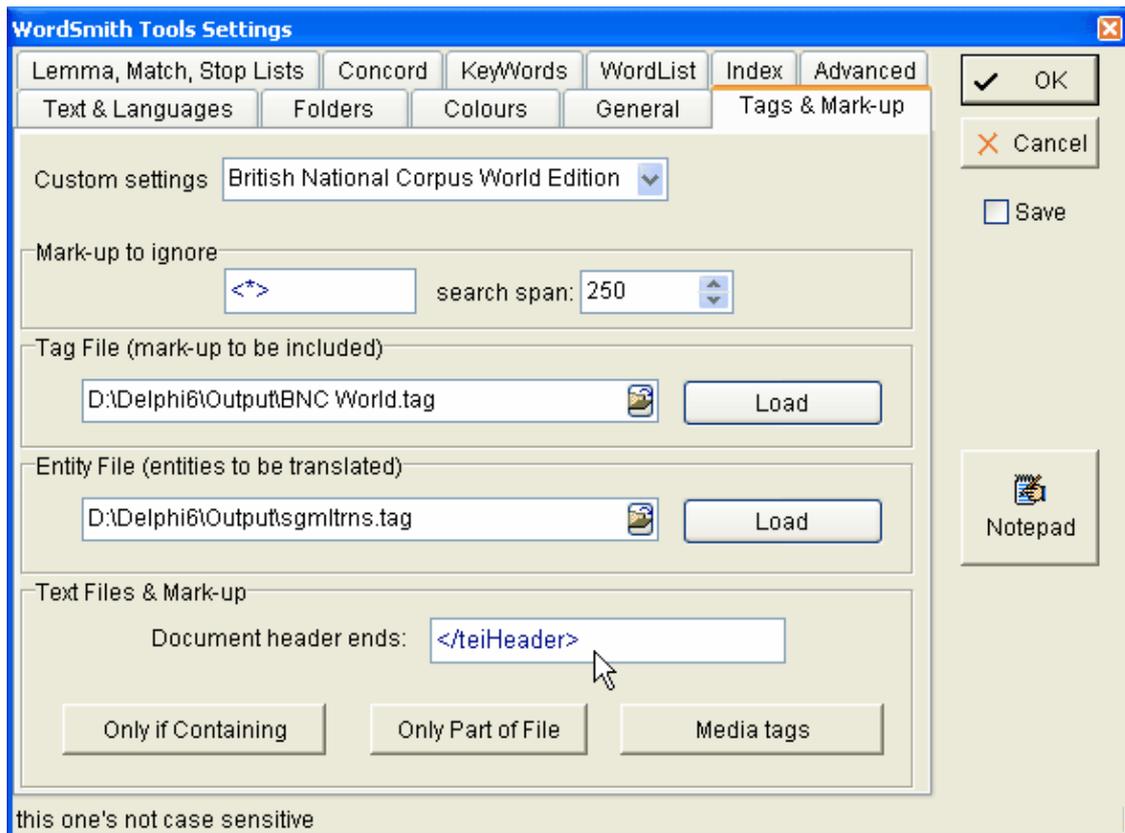
6.5 tags as selectors

Defaults

The defaults are: select *all* sections of *all* texts selected in [Choose Texts](#) but cut out all angle-bracketed tags.

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 *Settings | Adjust Settings | Tags | Only If Containing or Only Part of File*.



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.

Markup to ignore

If you want to cut out unwanted tags eg. in [HTML](#) 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.)

Tag File, Entity file

See [Making a Tag File](#).

Document Header

If you simply want to cut out a document header (a repeated header containing copyright notices as is found at the start of every BNC text), you just 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.)

For more complex searches, you might want to choose the [Only If Containing](#) or [Only Part of File](#) 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](#) to check whether it contains the words you've specified;
2. Cut out any section specified as a "[section to cut](#)";
3. If there are "[sections to keep](#)", cut out everything which is not within them;
4. Cut [start of each line](#), if applicable;
5. Process any entity references you want to [translate](#);
6. [Ignore](#) any tags not to be retained (see the "Mark-up to ignore" section of the screenshot above).

See also: [Overview of Tags](#), [Making a Tag File](#), [Tag Handling](#), [Tag Concordancing](#), [Showing Nearest Tags in Concord](#), [Viewing the Tags](#), [Types of Tag](#)

6.6 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](#) of the file) to check that it meets your requirements.

You can specify up to 12 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](#) and can save the list for later use as [favourites](#).

Examples:

You only want text files containing both the word `cats` and the word `dogs`: type `cats` into the first box, and `dogs` below it.

You want either `roses` or `violets`, and `flowers` must be present too: write `roses` and `violets` into the first two boxes, beside each other. Write `flowers` in the leftmost box on the next row down.

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](#), [Selecting within texts](#), [Filtering your text files](#) using [Text Converter](#).

6.7 selecting within texts

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 Cut

If you are using text files with [SGML](#), [XML](#) or [HTML](#) headers (e.g. the British National Corpus) you may simply want to cut out the header from your word lists, concordances, etc. as shown in the [Document header example](#).

For more complex choices, you may here specify what is to be cut, where it starts (for example <HEAD>) and where you want to cut to (e.g. </HEAD>). You can choose to cut out up to 3 different and separate sections (<HEAD> to </HEAD> or <BODY> to </BODY>). This function cuts out any section located as many times as it is found within the whole text.

Sections to Keep (contexts)

You want to select one section of a 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 **Peter:** (these would get all of Mary's contributions in the discourse but nothing else).

Naturally you must be sure that there is something unique like a < or > symbol to define each section. For example, in the case of **Mary:** and **Peter:** you'd want to be sure that every contribution made by Mary has a colon immediately following her name, and that all her contributions were followed by **Peter:**. This function is case sensitive (so it would not find **MARY:**).

If you used <H1> to </H1> with this function in [HTML](#) text you'd get all the major headings in your texts, however many, but nothing else.

You can choose to use 2 different sections, e.g. <Intro> to </Intro> to get the introduction and <Conclusion> to </Conclusion> to get the conclusion as well. 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>.

See also: [Tags as Selectors](#), [Only if containing <x>](#).

6.8 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](#).

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 ...

Tags to retain (mark-up to be included)

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 \wsmith4 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](#) (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" (excusez-moi, les gars).

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](#). 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](#).

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"
```

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.



Entity File (entities to be translated)

A tag file for translation of one entity reference into another uses the following syntax: entity reference to be found + space + replacement. For example:

```
&Eacute; É
&eacute; é
```

A sample tag file for translation (`\wsmith4\sgmltrns.tag`) is included with your installation: you could make a new one by reading it in, altering it, and saving it under a new name.

See also: [Overview of Tags](#), [Handling Tags](#), [Showing Nearest Tags in Concord](#), [Tag Concordancing](#), [Types of Tag](#), [Viewing the Tags](#), [Using Tags as Text Selectors](#)

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.

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 full stops or question marks or exclamation marks followed by a capital letter.

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](#), [Handling Tags](#), [Showing Nearest Tags in Concord](#), [Tag Concordancing](#), [Types of Tag](#), [Viewing the Tags](#), [Using Tags as Text Selectors](#)

WordSmith Tools

Concord

Section

VIII

7 Concord

7.1 purpose



Concord is a program which makes a [concordance](#) using [DOS](#), [Text Only](#), [ASCII](#) or [ANSI](#) 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, 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](#) them out, then [printing](#).

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.

7.2 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.

7.3 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 ...

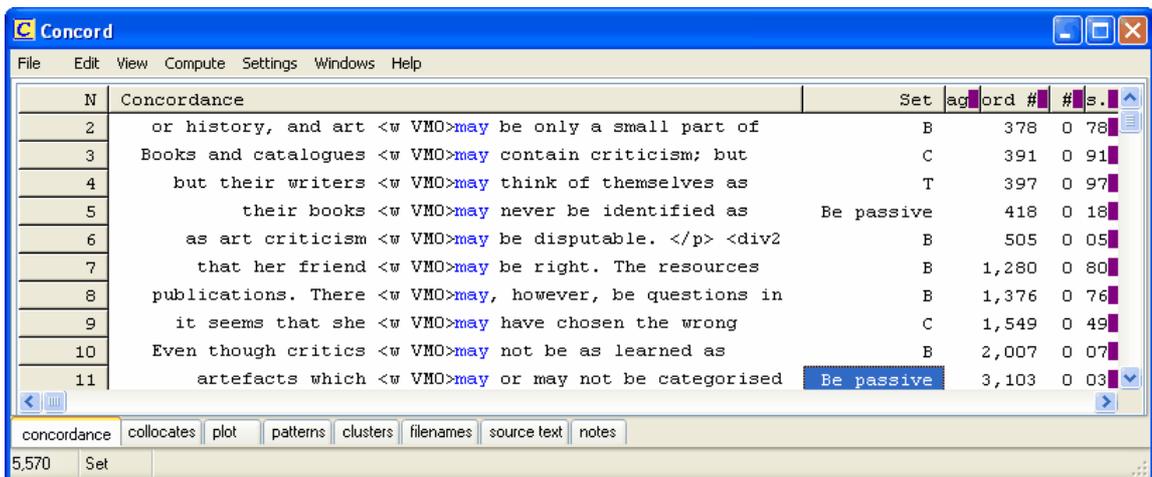
```

Concord will give equal space to the blanks so that the size of the blank doesn't give the game away.

7.4 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.



Here the user has used **B** where the verb following **may** is the verb **BE** but has also distinguished between **BE** as main verb and AUX **BE** in passive constructions, other verbs being classified according to their initial letter.

If you simply press a letter or number key while the [edit v. set v. type-in](#) mode is on Set (as it is in the screenshot above) you will get the concordance line marked with that letter or number in the Set column.

If you want to type something longer, double-click the set column and you'll get a chance to type more.

To correct a mistake, press the space key.

You can later [sort](#) the concordance lines using these categories as shown here, simply by clicking on the header *Set*.

The screenshot shows the Concord software window with a menu bar (File, Edit, View, Compute, Settings, Windows, Help) and a toolbar. The main window displays a concordance table with the following columns: N, Concordance, Set, ag, ord #, #, s. The table contains 10 rows of results for the search term 'may'. Below the table is a navigation bar with buttons for 'concordance', 'collocates', 'plot', 'patterns', 'clusters', 'filenames', 'source text', and 'notes'. At the bottom, there is a status bar showing '5,570 Set B'.

N	Concordance	Set	ag	ord #	#	s.
1	but their writers <w VMO>may think of themselves as	T		397	0	97
2	Books and catalogues <w VMO>may contain criticism; but	C		391	0	91
3	it seems that she <w VMO>may have chosen the wrong	C		1,549	0	49
4	their books <w VMO>may never be identified as	Be passive		418	0	18
5	artefacts which <w VMO>may or may not be categorised	Be passive		3,103	0	03
6	their main topics <w VMO>may be personalities or	B		371	0	71
7	or history, and art <w VMO>may be only a small part of	B		378	0	78
8	as art criticism <w VMO>may be disputable. </p> <div2	B		505	0	05
9	that her friend <w VMO>may be right. The resources	B		1,280	0	80
10	publications. There <w VMO>may, however, be questions in	B		1,376	0	76

See also : [edit v. type-in](#) mode.

7.5 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](#). The [default](#) is 5,5 (5 to left and 5 to right) but you can go up to 25 on either side.

To set collocation horizons and other **Concord** settings, in the main **WordSmith Controller** menu at the top, choose *Adjust Settings*, then *Concord*.

See also: [Collocate Settings](#)

7.6 collocate settings

To set collocation horizons and other **Concord** settings, in the main **WordSmith Controller** menu at the top, choose *Adjust Settings*, then *Concord*.

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 **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](#).

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).

7.7 collocates display

The point of it...

The point of all this is to work out characteristic lexical patterns. 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. Collocational linkages which involve grammatical items are often referred to as *colligation*.

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](#) 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](#).

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 **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

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.

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

The frequency display can be [re-sorted](#) (🔄) and you can recalculate the collocates (📊) if you [zap](#) entries from the concordance or change the [horizons](#).

See also: [Collocation](#), [Collocation Relationship](#), [Mutual Information](#)

7.8 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.

Choose a wordlist

The collocation display initially shows the collocates without any relation information (eg. 0.00); this is because to compute the relationship we need to know

- a) how often each collocate appears in the corpus we're using,
- b) how often the [search-word](#) appears in the corpus, and
- c) how often they come together within the [horizons](#) selected.

The problem is that although b) and c) are known at the time the concordance is computed, a) is not known without doing a concordance or wordlist for each collocate....

So you are asked to choose an appropriate wordlist (created by WordSmith of course!), which will know the frequencies for each word. It's up to you to choose a wordlist which actually relates to the concordance you've done!

Type of Relation

Choose in the main [Controller Concord settings](#) which type of relation you wish to compute.

See also: [Collocation](#), [Collocate display](#), [Mutual Information](#)

7.9 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*.

The point of it...

By examining the collocates you can find out more about "the company the word keeps", which helps to show its meaning and its usage.

Options

You may compute a concordance with or without collocates: 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](#).

You can re-compute collocates after editing your concordance.

If you want to filter your collocate list, use a [match list](#) or [stop list](#).

[Re-sort](#) a collocate list in a variety of ways.

You can see the [strength of relationship](#) between the word and the search-word which the concordance was based on.

Collocates can be [viewed](#) after the concordance has been computed.

Technical Note

The [literature](#) 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](#), which are a pointer to "coherence collocates".

Concord will supply "neighbourhood collocates". **WordList** allows you also to study [Mutual](#)

[Information](#).

See also: [collocation display](#), [collocation settings](#), [collocation relationship](#), [mutual information display](#).

7.10 Concord: clusters

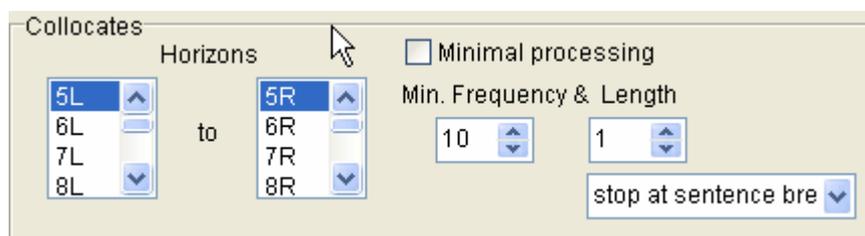
The point of it...

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](#).

How it does it...

Clusters are computed automatically if this is not disabled in the main [Controller](#) settings for Concord (*Adjust Settings | Concord*) where you will see something like this:



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.)

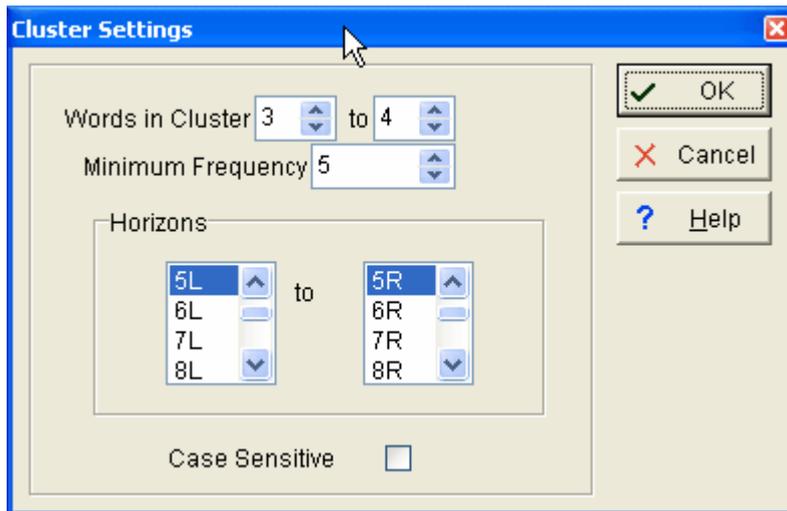
Clusters are sought within these limits: default: 5 words left and right of the search word, but up to 25 left and 25 right allowed. The default is for clusters to be three words in length and you can choose how many of each must be found for the results to be worth displaying (say 3 as a minimum frequency).

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](#), the idea of "stop at sentence breaks" (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.



When you press OK, clusters will be computed. In this case we have asked for 3- or 4-word clusters and get results like this:



Most of the clusters are 3-word clusters but the highlighted one has 4 words here. Line 10 shows there are 6 cases of `if this is`, and line 12 shows that most of them are `if this is a`.

It's a dependent window

Each Cluster window is dependent on the Concordance from which it was derived. If you close the original concordance down, they will disappear.

See also: [general information on clusters](#), [WordList Clusters](#).

7.11 Concord: dispersion

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.

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](#)

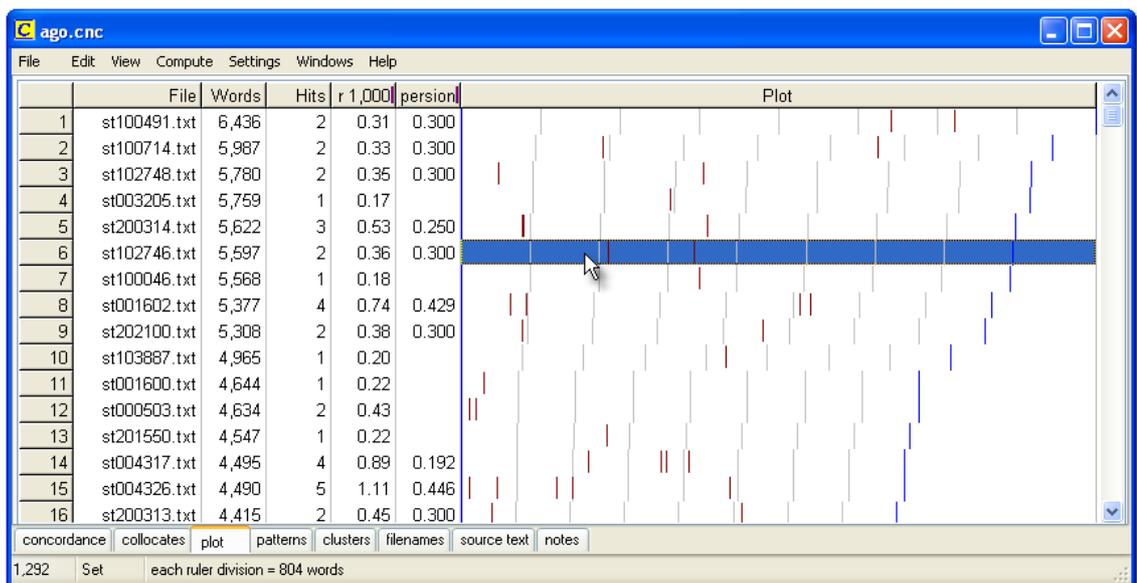
Plot a plot showing where they cropped up, where the left edge of the plot represents "Once upon a time" and the right edge is "happily ever after".



Here we see a plot of "O" and another of "AH" from the play Romeo and Juliet. They are on 2 separate lines because there were 2 search-words. There are more "O" exclamations than "AH"s. There is a "ruler" splitting the display into 8 segments, and the status bar tells us each segment represents about 3150 running words of the play.

The plot is initially [sorted](#) by no. of words per 1,000.

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.



The screenshot shows "uniform plot" -- as the statusbar says, each ruler segment represents 800 words in these dispersion plots of "ago". If you look at the Words column, you will see that the number of words in each file varies, which is why the blue right plot edge and the ruler marks vary in position.

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](#) and can get your spreadsheet to make graphs etc, as [explained here](#).

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 because the data come from a concordance which you should [Save](#), or *Print to File*. You can *Copy* to the [clipboard](#) (Ctrl-Ins) and then put it into a word processor as a graphic, using Paste Special.

See also: [plot and ruler colours](#), [plot dispersion value](#).

7.12 Concord: 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](#) 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](#).

7.13 Concord: viewing options

These menu options toggle on and off. When on, they're checked. They include:

Sentence Only

This will show only the sentence in which the search-word appears.

Tags and Spaces Cut

If you have specified any [tags to retain](#), these will normally be visible in your concordance. If you wish to hide them, toggle this menu option. The same option will also cut out any redundant spaces in your concordance line; these might be caused by the presence of [tags](#) which have been ignored.

See also: [showing nearest tags](#), [Blanking out](#) the search-word.

7.14 Concord: handling 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. To play the media file, press Control/M or choose *File | Play media file*. The tag will be [parsed](#) to identify the file needed, if necessary downloading it from a web address, and then played.

See also: [Handling Tags](#), [Making a Tag File](#), [Showing Nearest Tags in Concord](#), [Tag Concordancing](#), [Types of Tag](#), [Viewing the Tags](#), [Using Tags as Text Selectors](#), [Tags in WordList](#)

7.15 Concord: what you see and do

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. Click for information on the [menu](#).

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), source text [filename](#), and how far into the file it comes (as a percentage). See below for an explanation of the **purple blobs**.

Set

This is where you can classify the entries yourself, using any letter, into [user-defined categories](#). 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](#), 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](#).

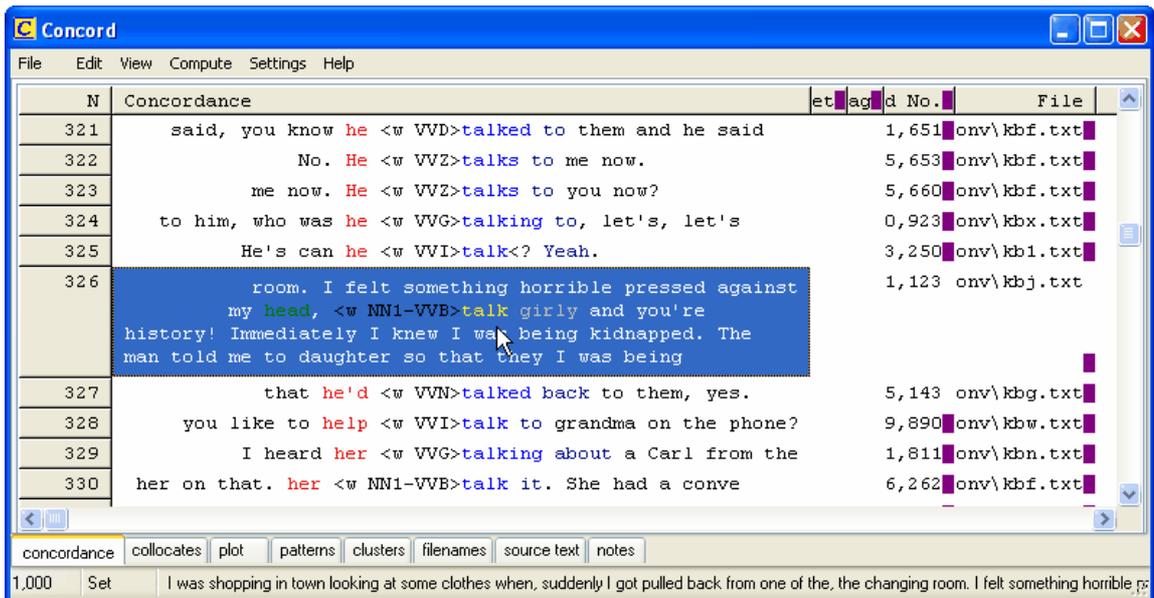
Stretching the display to see more

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.)



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

Viewing the original file (📄)

Available if it is still on the disk where it was when the concordance was originally created.

Status bar

The [status bar](#) panels show

- the number of entries (1,000 in the screenshot)
- whether we're in "Set" or "Edit" mode;
- the current concordance line from its start.

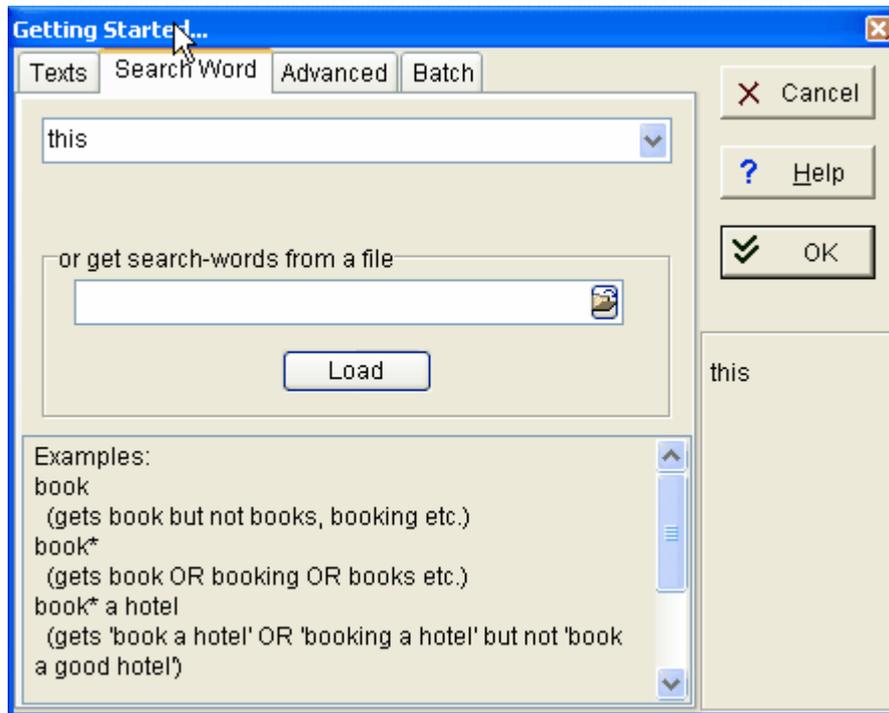
See also:

- [Re-sorting](#) your concordance lines
- [User-defined categories](#)
- [Altering the View](#)
- [Blanking out](#) the search-word
- [Collocation](#) (words in the neighbourhood of the search-word)
- [Plot](#) (plots where the search-word came in the texts)
- [Clusters](#) (groups of words in your concordance)
- [Text segments in Concord](#)
- [Editing the concordance](#)
- [Zapping entries](#)
- [Saving and printing](#)

7.16 concordance settings

Search Word or Phrase and/or Tag

Type the [word or phrase](#) Concord will search for when making the concordance, or the name of a [file of search words](#). You may also choose from a [history list](#) of your previous search words. For details of syntax, see [Search Word Syntax](#) or the set of examples shown in this screenshot:

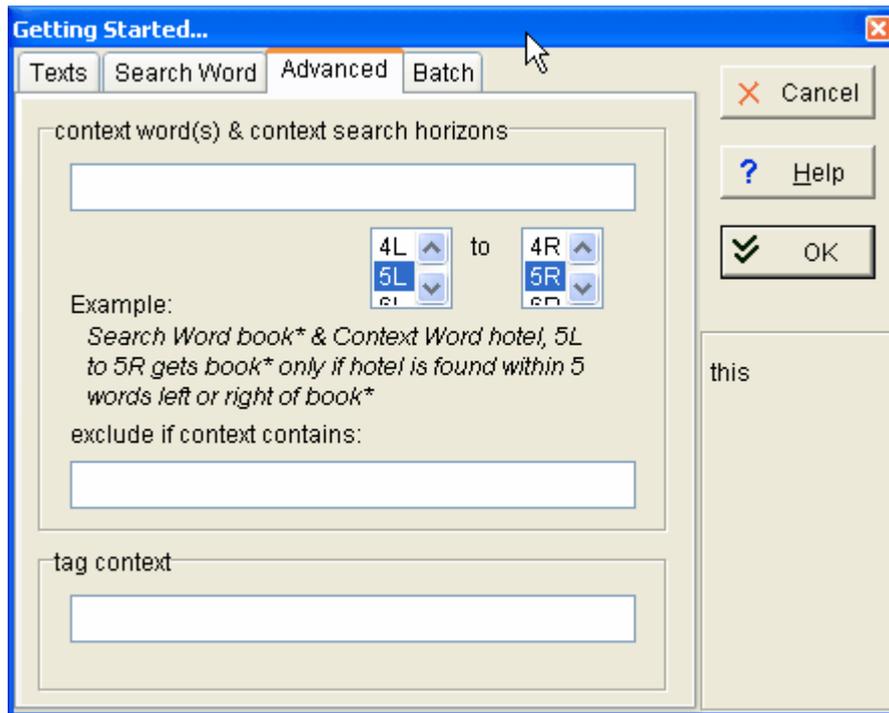


If you want to do many concordances in a [file-based search](#), first prepare a small text file containing the search words, e.g. containing

```
this
that
the other
==Major*==
```

Press the file button to locate your text file, then 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



Context Word(s) and Context 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](#). Or if the search word is `book*` and the *exclusion* word is `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 `book*` and the exclusion word is `booked`, you'll get `book`, `books`, `booking`, `bookable`, but not `booked`.

Tag Context

The tag context is the context of tags as defined in your [tag-file](#), to the left of the search-word. In this example, the tag context is `<u speaker=Simon>`:

```
.....<u speaker=Simon> Between you and me, I wish I hadn't booked
that hotel ....
```

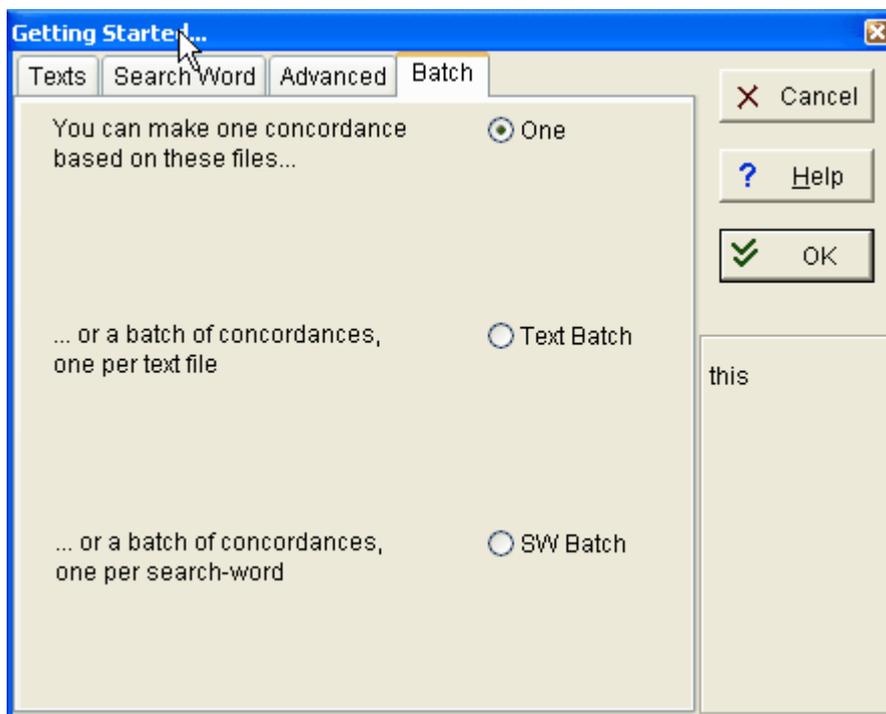
The tag context actually can include a bit more than that -- if this chunk is in `<section 1>`, and there is a tag showing the end of section 1 (probably `</section>`) which occurs anywhere to the right of the search-word, then the full tag context might be `<section 1><u speaker=Simon>` or even more. So a tag context is actually the whole set of tags which are still in operation at the point where your search-word was found.

You may wish to find a word or phrase depending on the tag context. In that case you specify tag attributes which you require.

For example, if the search word is `book*` and the tag context is `Simon`, you'll get `book`, `books`, `booking`, `bookable`, `booked` only if `Simon` is found in the tag context.

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](#) which can be viewed separately (*Text Batch*). If you have multiple search-words in a [file-based search](#) as explained above, you may want each result saved separately (*SW Batch*).



Other settings affecting a concordance are available too: see [WordSmith Controller Concordance Settings](#); [Accented characters](#); [Choosing Language](#), [Context Word\(s\) & Context Search Horizons](#)

7.17 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](#).

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](#).

It doesn't matter whether you are using a [tag file](#) 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](#) on either side.

Are you concordancing on tags?

If you are asked this and you are concordancing 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](#).

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).

See also: [Overview of Tags](#), [Handling Tags](#), [Showing Nearest Tags in Concord](#), [Search word or phrase](#), [Types of Tag](#), [Viewing the Tags](#), [Using Tags as Text Selectors](#)

7.18 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](#) 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.

If you have specified a context word, you can re-sort on it. Also, the context words will be in their own special [colour](#).

Syntax is like that of the [search word or phrase](#),

* 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. "*"

7.19 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 (**Alt-Z** or ) to [zap](#) the deleted lines. If you're a teacher you may want to [blank](#) out the search words: to do so, press the spacebar. Pressing the spacebar again will restore it, so don't worry!

See also: [Window Management](#)

7.20 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](#).

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 (`\wsmith4\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.

Lemmas and file-based concordancing

Note that where Concord has been [called up](#) from WordList, and the highlighted word in the word list is the head entry with [lemmas](#), 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 `\wsmith4` 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 searchwords, 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](#).

7.21 nearest tag

Concord allows you to see the nearest tag, if you have specified a [tag file](#), which teaches Oxford 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](#) 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>
```

You can [sort](#) on the nearest tags. In the shot below, a concordance of **such** has been computed using BNC World 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	<w 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	5
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	5
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	<w TOO>to <w VVI>make <w DT0>such <w NN2>identifications,		<w DT>	2,792	08	13
25	<w CJC>or <w DT0>such <w NN2>questions <w		<w DT>	3,133	20	16
26	<w CJC>and <w DT0>such <w NN2>treatises <w		<w DT>	3,170	21	35
27	<w VMO>may <w VVI>take <w DT0>such <w ATO>a <w		<w DT>	3,570	36	6
28	<w NN1>creation of <w DT0>such <w NN2>theories <w		<w DT>	3,722	41	25
29	<w PRP>by <w DT0>such <w ATO>an <w		<w DT>	4,857	95	17
30	<w PRP>in <w DT0>such <w NN2>papers <w CJS>as		<w DT>	4,951	99	5
31	<w NN1>criticism of <w DT0>such <w NN1>art <w PRP>within		<w DT>	7,534	01	5
32	<w DT0>This <w VVD>took <w DT0>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](#) have the

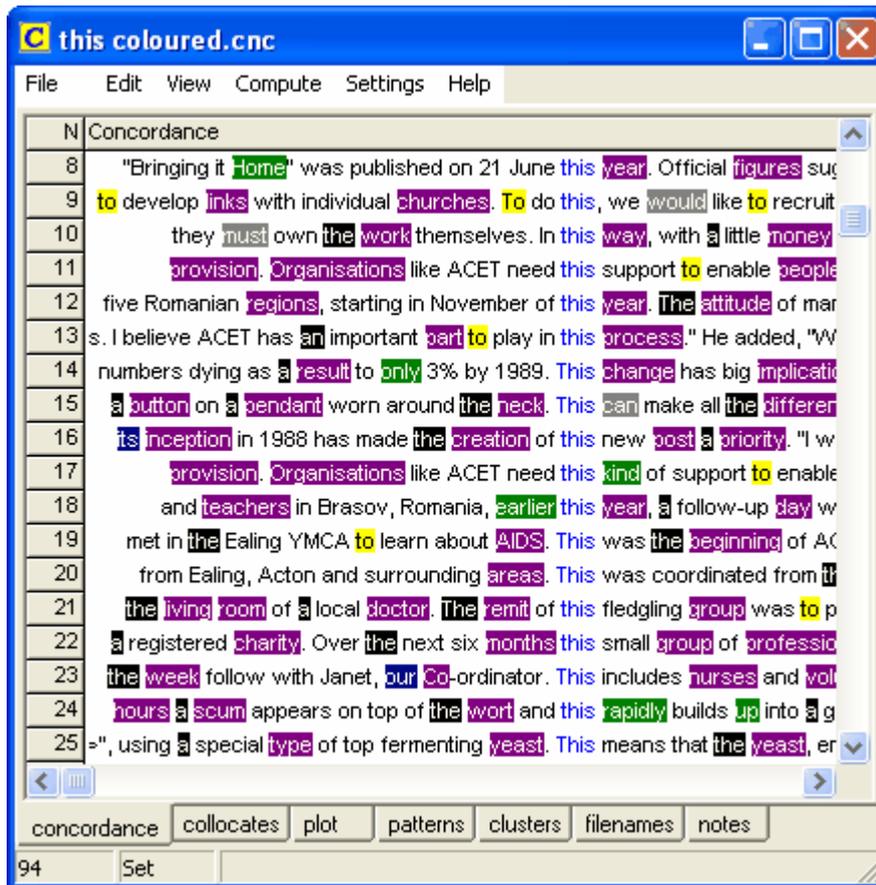
same format. For example, if Tags to Ignore has <*>, any tags such as <title>, <quote>, etc. will be cut out of the concordance unless you specify them in a [tag file](#). If so, specify the tag file and run the concordance again.

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 World Edition text with the tags in colour:

N	Concordance
3	<wNN1>availability. <wDTO>This <wVBZ>is <wAVD>so <wPNP>we
4	<wNN1>spread <wPRF>of <wDTO>this <wAJD>terrible <wNN1>disease
5	<wNN1>need <wPRP>for <wDTO>this <wNN1>service. <wNN1>Home
6	<wNN1>opening <wPRF>of <wDTO>this <wNN1>office <wPRP>in
7	<wCJS>even though <wDTO>this <wVBZ>is <wAVD>increasingly
8	<wPRP>on 21 <wNPO>June <wDTO>this <wNN1>year. <wAJD>Official
9	<wTOO>To <wVDI>do <wDTO>this, <wPNP>we <wVMD>would
10	<wPRP>in <wDTO>this <wNN1>way, <wPRP>with <wATO>
11	<wVVB-NN1>need <wDTO>this <wVVI>support <wTOO>to
12	<wNPO>November <wPRF>of <wDTO>this <wNN1>year. <wATO>The
13	<wVVI>play <wPRP-AVP>in <wDTO>this <wNN1>process. &equo; <wPNP>He
14	3&percent; <wPRP>by 1989. <wDTO>This <wNN1>change <wVHZ>has
15	<wATO>the <wNN1>neck. <wDTO>This <wVMD>can <wVVI>make
16	<wNN1>creation <wPRF>of <wDTO>this <wAJD>new <wNN1>post <wATO>sh

and here is a view showing the same data, with *View | Hide Tags* selected.





The tags themselves are no longer visible, and only 6 types of tag have been chosen to be viewed in colour.

See also: [Overview of Tags](#), [Handling Tags](#), [Making a Tag File](#), [Tagged Texts](#), [Types of Tag](#), [Viewing the Tags](#), [Using Tags as Text Selectors](#)

7.22 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.

The minimum frequency for one of the words to be shown at all, is the [minimum frequency for collocates](#).

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.

7.23 re-sorting

How to do it...

Sorting can be done simply by pressing the top row of any list. Or by pressing F6 / Ctrl/F6. Or by choosing the menu option.

As a concordance is generated, it will appear in the order in the text file(s) which the concordance came from: file order. As soon as the concordance is completed, it will then get re-sorted in the order set in the current [defaults](#). If you don't want re-sorting to occur, set the defaults to file,file.

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](#), the distance to the nearest tag, a [set category](#) of your own choice, or original file order (file).

Main Sort

The listing can be sorted by three criteria at once. A Main Sort on Left 1 will sort the entries according to the alphabetical order of the word immediately to the left of the search word. A second sort (... then by ...) on Right 2 would re-order the listing by tie-breaking, that is: only where the Left 1 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 (...finally by) 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 first and second sorts are the same (e.g. Left 1 and Left 1) or if the "activated" box is not checked.

sorting by set (user-defined categories)

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**. The sort will put the classified lines first, in category order, followed by any unclassified lines (which will appear in a light grey colour). See [Nearest Tag](#) for details of sorting by tags.

The colour of the search word will change according to the sort system used.

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](#), [KeyWords sort](#), [Choosing Language](#)

7.24 re-sorting: collocates

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, from 25 words to the left of the search word to 25 words to the right. Just press the header of a column to sort it. Press again to toggle the sort between ascending to descending.

The point of it...

is to find patterns of collocation, so as to more fully understand the company your search-word keeps.

The choices depend on the [collocation horizons](#).

See also: [Collocation](#), [Collocation Display](#)

7.25 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](#)

7.26 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

concordance | collocates | plot | patterns | clusters | filenames | source text | notes

100 | Set | iv1 n=TWO type=part> <head> <s n="1">THE CHURCH'S <w NN1>YEAR </head> <div2

See also: [Start and end of text segments.](#)

7.27 search word syntax

By default, Concord does a whole-word non-case-sensitive search.

Examples

search word	finds
book	<i>Book or book or BoOk</i>
book	<i>book, books, booking, booked</i>
*book	<i>textbook (but not textbooks)</i>
bo* in	<i>book in, books in, booking in (but not book into)</i>
book * hotel	<i>book a hotel, book the hotel, book my hotel</i>
bo* in*	<i>book in, books in, booking in, book into</i>
book?	<i>book, books, book; book.</i>
book^	<i>book, books</i>
b^^k	<i>book, back, bank, etc.</i>
==book==	<i>book (but not BOOK or Book)</i>
book/paperback	<i>book or paperback</i>

symbol	meaning	examples
*	disregard the end of the word, disregard a whole word	tele* *ness *happi*
book * hotel		Engl???
?	any single character (including punctuation) will match here	?50.00
#	any single number, 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) separates alternative search-words. You can specify alternatives within an 80-character overall limit	c:\text\frd.txt may/can/will
<>	beginning & end of tags	<w NN1>

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 "*"?"

Tags

You can also specify tags in your search-word if your text is tagged.

Examples:

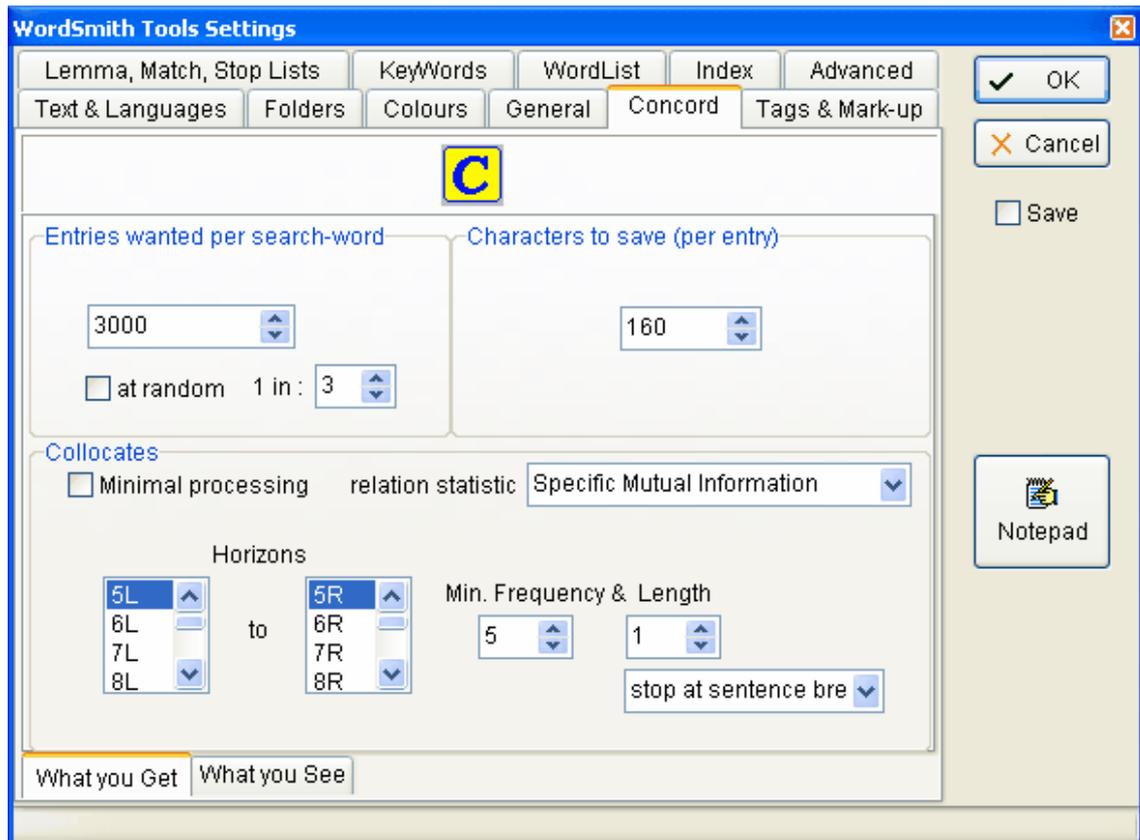
symbol	meaning	examples
<w NN1>*	single common noun (BNC)	<i>book, chair, elephant</i>
<w NN?>*	singular or plural common noun	<i>book, chairs</i>
<w NN1>t*	any single noun beginning with T or t	<i>table, teacher</i>
<w NN1>* <w NN1>*	two single common nouns in sequence	<i>campaign manager</i>

See also: [Tag Concordancing](#), [Context Word](#)

7.28 WordSmith controller: Concord: settings

These are found in the main [Controller](#) under *Adjust Settings | Concord*.

This is because some of the choices -- e.g. [collocation horizons](#) -- may affect other Tools.



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 below showing the options under *What you see*.

WHAT YOU GET

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**.

"*at random*" is a feature which allows you to randomise the search. Here **Concord** goes through the text files and gets the 100 entries by giving each hit a random one-in-three chance of being selected. To get 100 entries **Concord** will have found around 250-350 hits. You can set the randomiser anywhere from 1 in 2 to 1 in 1,000.

Characters to save

Here is where you set how many characters in a concordance line will be saved as text (as opposed to sending them to the [clipboard](#)). The default is 80 (minimum 20 and maximum 8,000). The reason for this is that you will probably want a fixed number of characters so that when using a non proportional font, such as Courier or Lucinda Console, the search-words line up nicely.

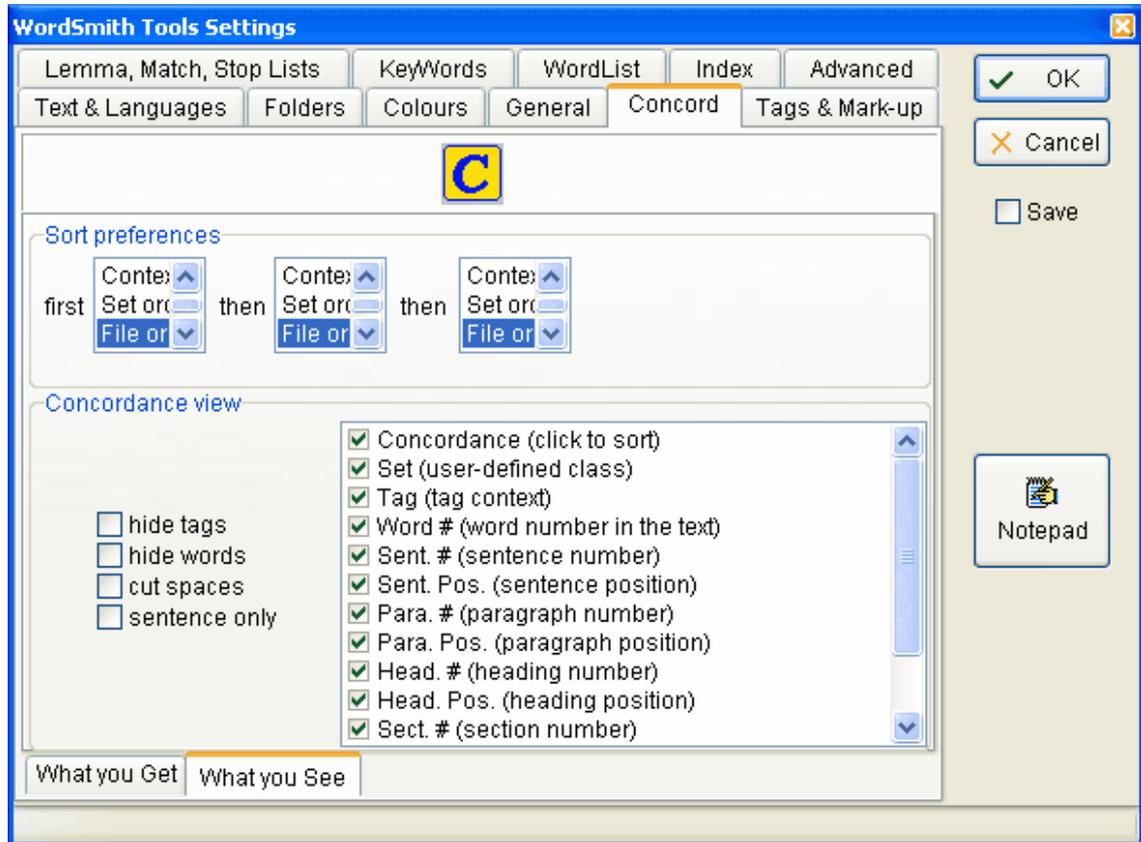
Collocates

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](#) or [Collocation](#)

Collocates relation statistic

Choose between Specific Mutual Information, MI3, Z Score, Log Likelihood. See [Mutual Information Display](#) for examples of how these can differ.

WHAT YOU SEE



Sort preferences

By default, **Concord** will sort a new concordance in original file order, but you can set this to different values if you like. For further details, see [Sorting a Concordance](#).

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](#).

See also: Concord Help Contents, [Collocation Settings](#).

WordSmith Tools

Key Words

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](#).

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](#) between them can be [plotted](#), made into a [database](#), and grouped according to their [associates](#).

8.2 Two word-list analysis

The usual kind of **KeyWords** analysis. It compares the one text file you're chiefly interested in, with a reference corpus based on a lot of text.

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.

See also [How Key Words are Calculated](#), [KeyWords Settings](#)

8.3 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](#) 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](#), [Associates](#), [Definitions](#), [Mutual Information](#)

8.4 associates

"Associates" is the name given to key-words associated with a [key key-word](#).

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.

You can set a minimum number of files for the association procedure, using *Settings / Database*.

See also: [definition of associate](#).

8.5 choosing files

Current Text Wordlist

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 wordlist(s) you're going to find the key words of.

Reference Corpus Wordlist

The the box below, you choose your [Reference Corpus](#) 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.

8.6 clumps

"Clumps" is the name given to groups of key-words [associated](#) with a [key key-word](#).

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](#), you may simply join those you think belong together

(by dragging), or regroup with help by pressing .

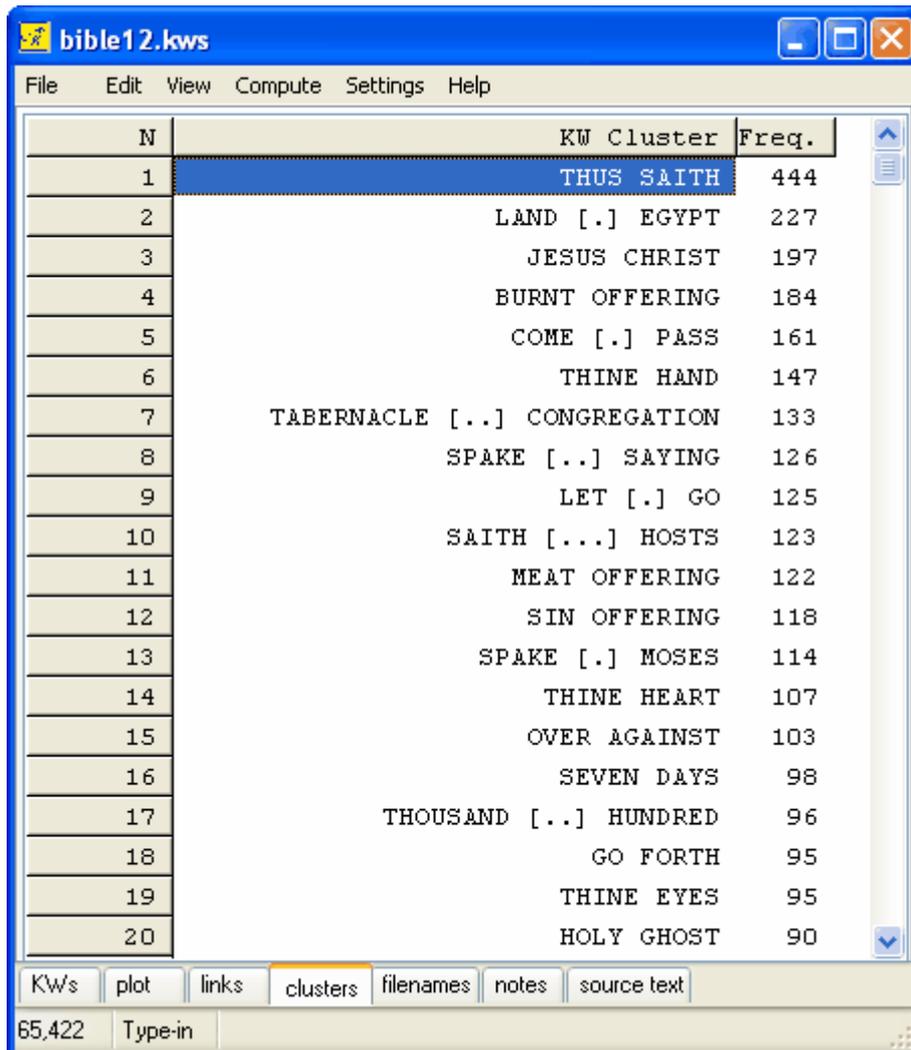
The listing shows clumps sorted in alphabetical order. You can re-sort by frequency (the number of times each key word in the clump appeared in all the files which comprise the clump).

See also: [definition of associate](#), [regrouping clumps](#)

8.7 KeyWords clusters

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.



N	KW Cluster	Freq.
1	THUS SAITH	444
2	LAND [.] EGYPT	227
3	JESUS CHRIST	197
4	BURNT OFFERING	184
5	COME [.] PASS	161
6	THINE HAND	147
7	TABERNACLE [...] CONGREGATION	133
8	SPAKE [...] SAYING	126
9	LET [.] GO	125
10	SAITH [...] HOSTS	123
11	MEAT OFFERING	122
12	SIN OFFERING	118
13	SPAKE [.] MOSES	114
14	THINE HEART	107
15	OVER AGAINST	103
16	SEVEN DAYS	98
17	THOUSAND [...] HUNDRED	96
18	GO FORTH	95
19	THINE EYES	95
20	HOLY GHOST	90

These are clusters computed using the Bible as source text. Each of the words here is "key" by comparison to a reference corpus; the clusters show cases where these KWs occur within the current collocation horizons. The [...] brackets represent cases where the KWs are not found together, e.g in *come [.] pass* there is one dot because the repeated occurrences are *come to pass*.

See also: [Plot calculation](#).

8.8 concordance

With a key word list on your screen, you can press **C** 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.

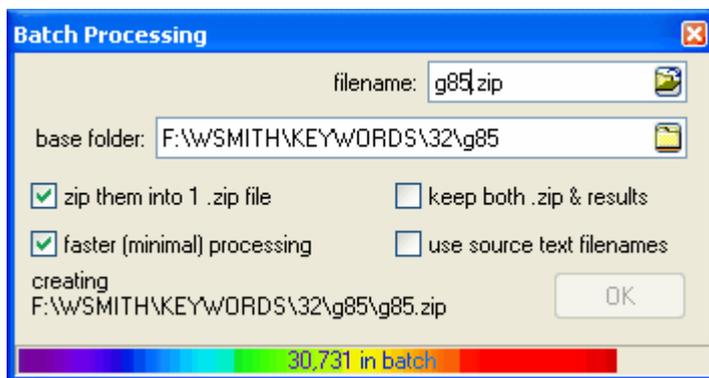
The point of it...

is to see these same key-words in their original contexts.

8.9 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](#) of word lists.
2. Use this to make a [batch](#) of keyword lists. Set "faster minimal processing" on as in this shot, so as to not waste time computing plots etc.



3. Now, in **KeyWords**, choose *New | Database*.

This enables you to choose the whole set of key word files.

Note that making a database means that only [positive](#) key words will be retained.

In the [Controller KeyWords settings](#) you can determine how many key-words there must be per text for your database.

Then specify where the resulting database is to be stored.

After you've made the database, you can then *Open* it to see the [key key-words](#) and [associates](#).

8.10 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.

8.11 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.

See also: [How Key Words are Calculated](#), [Definition of Key Word](#), [Creating a Database, Definitions](#)

8.12 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](#) is smaller than or equal to a [p value](#) 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 wordlists based on [word clusters](#).)

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 press **C**), to investigate why such words have cropped up with unusual frequencies.

See also: [How Key Words are Calculated](#), [Definition of Key Key-Word](#), [Definitions](#), [KeyWords Settings](#)

8.13 KeyWords database

(default file extension .KDB)

The point of it...

The point of this database is that it will allow you to see the "key-key-words" in your set of files. That is, the key-words which are most frequent 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.

By making up a database, you can sort these out. The ones at the top of the list, when you view them, will be those which are most typical of the genre. The list is ordered in terms of "key key-ness" but can be toggled 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 *Settings | KeyWords | Database*.

When viewing a database you will be able to investigate the [associates](#) 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](#), [Definition of key key-word](#)

8.14 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](#).
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. 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!

8.15 KeyWords: calculation

The "key words" are calculated by comparing the frequency of each word in the wordlist of the text you're interested in with the frequency of the same word in the reference wordlist. All words which appear in the smaller list are considered, unless they are in a [stop list](#).

If **the** occurs say, 5% of the time in the small wordlist 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 wordlist
the number of running words in the small wordlist
its frequency in the reference corpus
the number of running words 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](#) Log Likelihood test, which gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus.

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 wordlist.

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](#) and a plot.

8.16 KeyWords: links

The point of it...

is to find out which key-words are most closely related to a given key-word.

A [plot](#) 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.

What you see

This Links window shows the number of links followed by a column headed "in" and a percentage. This percentage represents the number of links divided by the total number of occurrences of the word in question (the "in" column number). Thus if you choose to see the links of *elephant*, and *elephant* crops up 10 times in your original text, and all 10 of those times it's found near the word *water*, (even though *water* occurs 40 times altogether), you'll see 100%. If you choose to see the links of *water*, the percentage next to *elephant* will be 25%.

The collocation [horizons](#) are those set in **Concord**, and go up to 25 words to left and right. The [default](#) is 5,5.

Double-click on any word in the [plot listing](#) to call up a window (up to maximum of 20 windows) which show the linked key-words.

See also: [Plot calculation](#), [KeyWords clusters](#)

8.17 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).

8.18 p value

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 words 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) so as to obtain fewer key words. Or you can set a low "maximum wanted" number in the main [Controller](#), under *Adjust Settings | KeyWords*.

If the [chi-square procedure](#) is used, the computed p value will only be shown if all appropriate statistical requirements are met (all expected values ≥ 5).

See also: [Definitions](#)

8.19 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](#).

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](#).

Note: this process depends on KeyWords being able to find the [source texts](#) which your original wordlist was based on.

You may find it useful to [export your plot](#) and make other graphs, as explained under [Save As](#).

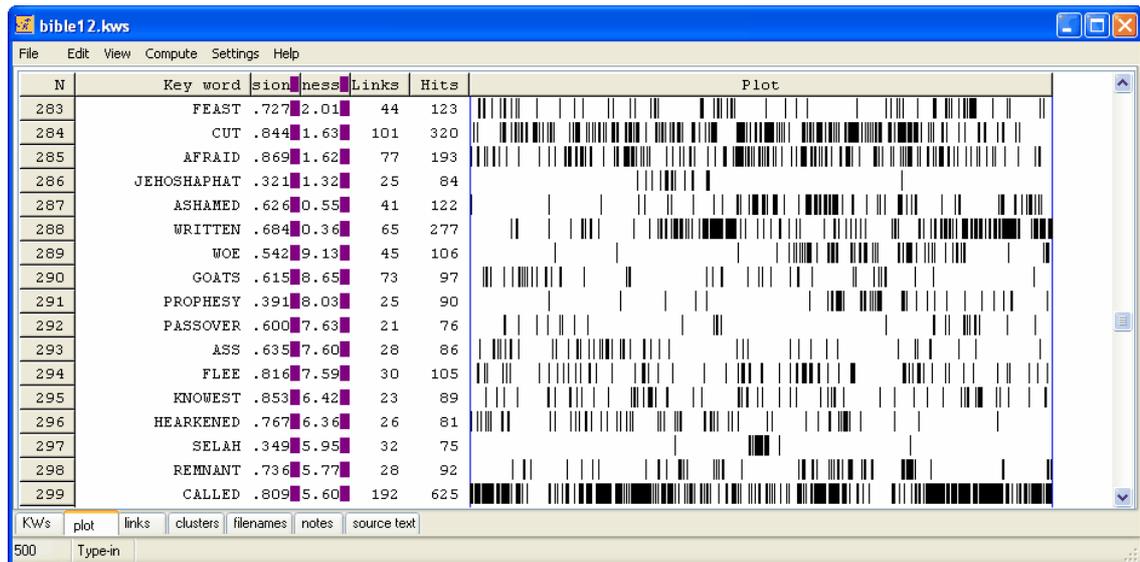
See also: [Plot Links](#), [Key words plot display](#)

8.20 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](#) 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 Bible, revealing where each term occurs. The

name **Jehoshaphat**, for example, occurs mainly about one third of the way through the text.



re-sorting

You can [re-sort](#) the listing using . Re-sorting rotates through the following types:

- first mention of each key word in the text
- dispersion within the text
- the original plot order (which is based on key-ness)
- alphabetical order
- total number of links with other key-words

links

This shows the total number of [links](#) between the key-word and other key-words in the same text, within the current collocation span ([default](#) = 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](#) 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](#) option because the plot has graphics, which cannot adequately be represented as text symbols, but you can *Copy* to the [clipboard](#) (Ctrl-Ins) 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](#) 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](#), [plot dispersion value](#)

8.21 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](#) (1997) shows how clumping reveals the different perceived roles of women in a set of *Guardian* features articles.

See also: [clumps](#)

8.22 re-sorting: KeyWords

How to do it...

Sorting can be done simply by pressing the top row of any list. Or by pressing F6 / Ctrl/F6. Or by choosing the menu option. Press again to toggle between ascending & descending sorts.

A **key words list** offers a choice between sorting by
 key-ness (the *keyest* 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 *keyest* 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](#) 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)

8.23 the key words screen

The display shows

1. each key word
2. its frequency in the source text(s) which these key words are key in, *italicised*.
3. the name of the source text file (or the word list file name if there's more than one) and %, also

in *italics*.

4. its frequency in the reference corpus

5. the name of the reference corpus file (or the corpus word list file name if that was based on more than one text) and %

6. keyness (chi-square or log likelihood [statistic](#))

7. [p value](#).

The calculation of how unusual the frequency is, is based on the [statistical procedure](#) 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](#) 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](#) (negative keywords), in a different colour.

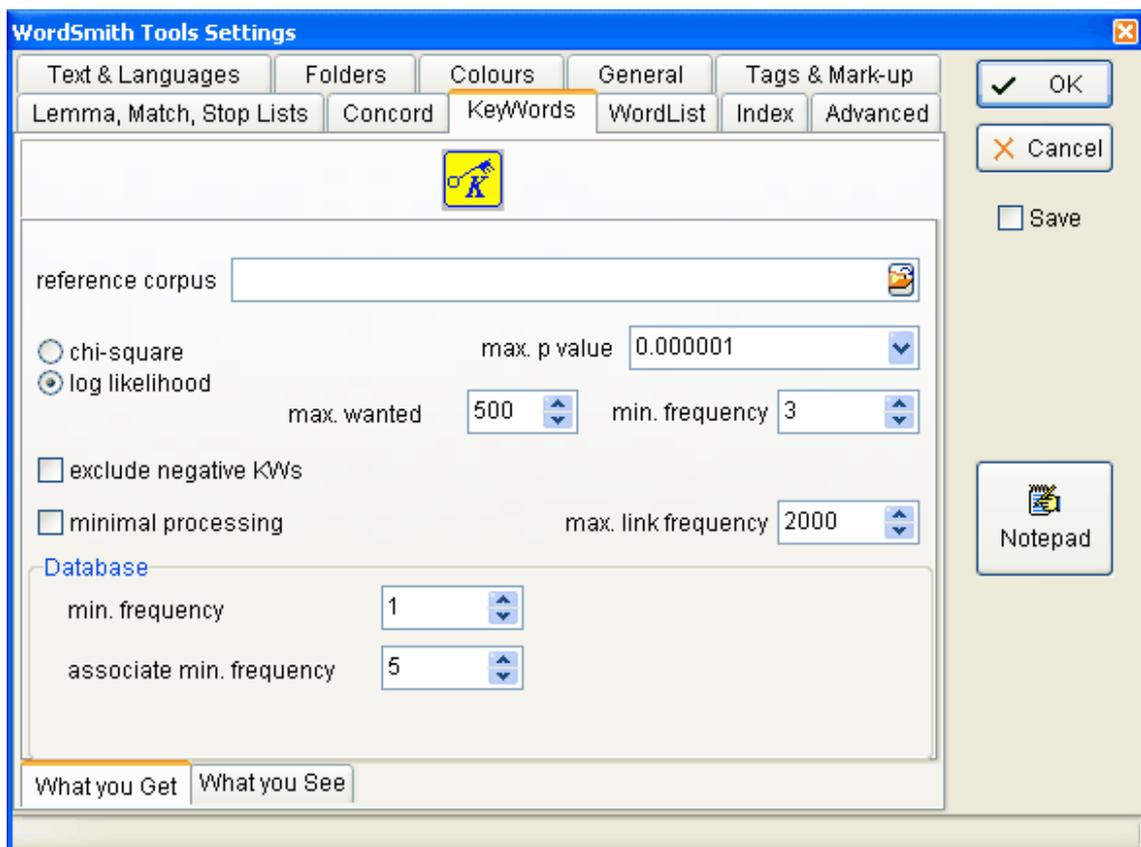
view button

This enables you to see the original source text using [Viewer](#), and will highlight the key words.

See  [layout](#) to change the individual colours or font of each column of data, e.g. if you don't like the italics.

8.24 WordSmith controller: KeyWords settings

These are found in the main [Controller](#) under *Adjust Settings | 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 [comparing word-lists](#).

Procedure

Chi-square or Log Likelihood. The default is Log Likelihood. See [procedure](#) for further details.

Max. p value

The default level of significance. See [p value](#) for more details.

Max. wanted (500) and Min. frequency (3)

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](#) 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.

Exclude negative KWs

If this is checked, KeyWords will not compute negative key words (ones which occur significantly *infrequently*).

Minimal processing

If this is checked, KeyWords will not compute [plots](#), [links](#) or [KW clusters](#) 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.

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](#). (It will still appear in the plots and list of KWs, of course.)

Database: minimum frequency

The default is 1. See [database](#).

Database: associate minimum frequency

The default is 5. See [associates](#).

See also: KeyWords Help Contents, [KeyWords calculation](#).

WordSmith Tools

WordList

Section

IX

9 WordList

9.1 purpose



This program generates word lists based on one or more [ASCII](#) or [ANSI](#) text files. The word lists are automatically generated in both alphabetical and frequency order, and optionally you can generate a [word index](#) 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](#);
- 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](#).

Within **WordList** you can compare two [lists](#), or carry out consistency analysis ([simple](#)) or ([detailed](#)) for stylistic comparison purposes.

These word-lists may also be used as input to the [KeyWords](#) 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".

See also: [WordList display](#)

9.2 auto-joining lemmas

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* heading). You will see all the joined entries at the top of

the list. It may be easier to [Unjoin](#) (Ctrl + F4) any mistakes than to confirm each one... Finally, sort on the *Word* and save.

See also: [Lemmatisation](#)

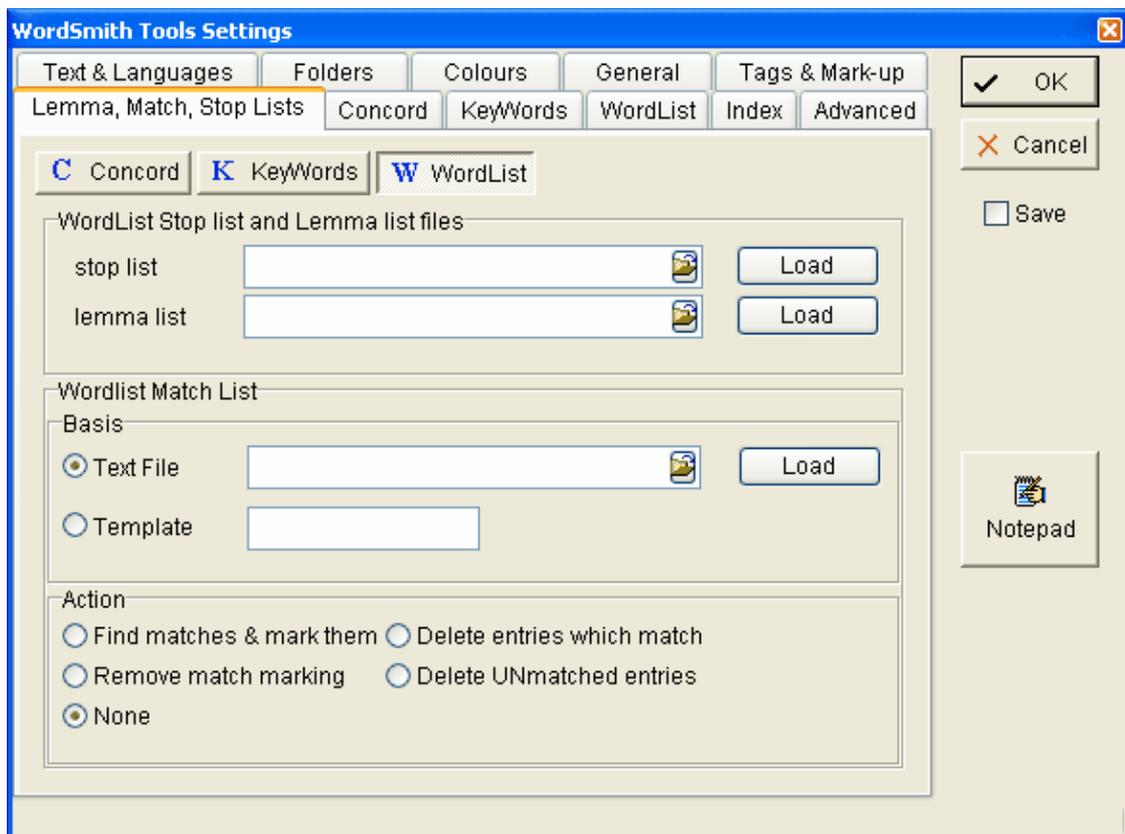
9.3 choosing lemma file

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](#).

How to do it

In the main Controller, *Settings | Adjust Settings | Lemma, Match, Stop lists*, you will see a screen like this:



Choose the appropriate button (for Concord, KeyWords or WordList) and type the file name or browse for 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
```

The **Controller** then reads the file and displays them (or a sample if the list is long).

The actual processing of the list only takes place when you choose the menu option *Match Lemmas* (≡) in WordList, Concord or KeyWords. See [Match List](#) for a more detailed explanation, with screenshots.

See also: [Lemmatisation](#), [Match List](#), [Stop List](#)

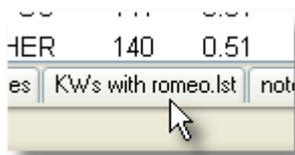
9.4 comparing wordlists

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

1. Open a wordlist.
2. In the menu, choose *File | Compare 2 wordlists*.
3. Choose a wordlist 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](#), *Adjust Settings*, *KeyWords* 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](#) from 0.1 to 0.000001 or what you will). The smaller the [p value](#), 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](#) format is similar to that used in [KeyWords](#).

See also: [Consistency Analysis](#), [Match List](#)

9.5 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 wordlist (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).

9.6 comparison display

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](#), and p [value](#).

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 wordlist. 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](#) program. The KeyWords analysis is slightly quicker and allows for batch processing.

The word **Lear** is the most key of all, it scores 304 on the keyness column. (It looks like 04.56 because the column hasn't been pulled any wider.)

N	Key word	Freq.	%	Freq.	RC.	%	yness	P	emmas
1	LEAR	229	0.83	0			04.56	00000	
2	KENT	173	0.63	0			29.91	00000	
3	FOOL	119	0.43	6	0.02		32.44	00000	
4	KING	70	0.25	2			77.51	00000	
5	EDMUND	52	0.19	0			69.00	00000	
6	GLOUCESTER	50	0.18	0			66.34	00000	
7	YOUR	223	0.81	103	0.40		53.21	00000	
8	EDGAR	37	0.13	0			49.08	00000	

The words above, in black, are key to Lear. Below, we see the middle of the listing --- the words in red are those which are key to Romeo. The word **most** is the last key word of Lear, and **death** the least key in Romeo; both have a keyness value of around 25 (positive or negative).

N	Key word	Freq.	%	Freq.	RC. %	yness	P	emmas
28	DAUGHTERS	28	0.10	2		25.34	04780	
29	TOM	19	0.07	0		25.20	05142	
30	HAVE	207	0.75	127	0.49	24.72	06596	
31	MOST	54	0.20	16	0.06	24.21	08608	
32	DEATH	22	0.08	71	0.27	24.29	08253	
33	BALTHASAR	0		19	0.07	25.20	05142	
34	GREGORY	0		19	0.07	25.20	05142	
35	MARRIED	0		19	0.07	25.20	05142	

Here at the bottom we see the words which are most key to the play *Romeo and Juliet*.

N	Key word	Freq.	%	Freq.	RC. %	yness	P	emmas
53	TYBALT	0		71	0.27	94.23	00000	
54	BENVOLIO	0		80	0.31	06.19	00000	
55	MERCUTIO	0		83	0.32	10.17	00000	
56	FRIAR	0		100	0.38	32.77	00000	
57	NURSE	1		149	0.57	87.38	00000	
58	CAPULET	0		145	0.56	92.63	00000	
59	JULIET	0		178	0.68	36.57	00000	
60	ROMEO	0		296	1.14	94.02	00000	

The word which is most outstanding (key) here is Romeo, with a keyness score of 394 (the column needs to be pulled wider).

9.7 consistency analysis (detailed)

This function does exactly the same thing as [simple consistency](#), 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 wordlists which you have called up. The display will order the words, so that the first group contains all those which occur in all versions, then those which come in all versions but one, and so on down to those which occur in only one version.

N	Word	Freq.	Texts	s	t	00.txt	0a.txt	0w.txt	01.txt	1d.txt	1e.txt	1f.txt
1	A	1,151	7	135	174	141	206	116	214	165		
2	ABOUT	77	7	13	5	5	16	7	20	11		
3	ACT	16	7	1	2	1	1	7	1	3		
4	AFTER	53	7	5	8	13	5	3	13	6		
5	AGAIN	36	7	4	1	12	2	2	9	6		
6	ALL	120	7	14	9	34	23	7	11	22		
7	ALSO	89	7	16	9	21	11	5	17	10		

Within each set the words are ordered alphabetically. The Freq. 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 **about** occurred in all 7 texts, it occurred 77 times in all, and it was most frequent in **1e.txt** at 20 occurrences. Statistics and filenames can be seen for the set of 7 texts used here by clicking on the tabs at the bottom. Notes can be edited and saved along with the detailed consistency list.

Note that the filename is test.**dcl** (**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.

How to do it...

In the window you see when you press **New...** () you will be offered a tab showing detailed consistency. Choose your word-lists and press **compute Detailed Consistency now**.

Each column can be sorted by clicking on its header column (**Word**, **Freq.** etc.). To get the words which occurred in all 7 texts to the top, I clicked **Texts**.

See also: [Consistency Analysis \(Simple\)](#), [Comparison Display](#), [Comparing Word-lists](#), [Match List](#), [Column Totals](#)

9.8 consistency analysis (simple)

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](#) 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 \(Detailed\)](#), [Comparing Word-lists](#), [Match List](#)

9.9 lemmas

You may want to store several entries together: e.g. `want`; `wants`; `wanting`; `wanted` as members of the same lemma.

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`.

Both the alphabetical and the frequency lists will be correctly updated, though the frequency list may not reflect the true order until after the file has been re-ordered by [zapping entries](#). 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.

N	Word	Freq.	%	Texts	%	Lemmas	Set
23	A GAME OF	5		4	25.00		
24	A GOD UNKNOWN	6		1	6.25		
25	A GOOD DEAL	15		1	6.25	a good deal[5] a great deal[10]	
26	A GREAT DEAL	10		5	34.25		

frequency alphabetical statistics filenames notes

3,554 Type-in A GREAT DEAL

Here we see a word list based on [3-word clusters](#) where originally `a good deal` had a frequency of 5, but has been joined to `a great deal` and thereby gained 10.

If you cannot see all the items you want to join in one screen, you can do the same thing using function keys.

1. Use 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 F5) will show they're marked too, in white. (If you change your mind you can press F5 again 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 first (the head).

To Un-join

If you select an item which has lemmas visible at the right and press Control/F4, this will unjoin the entries.

File-based joining

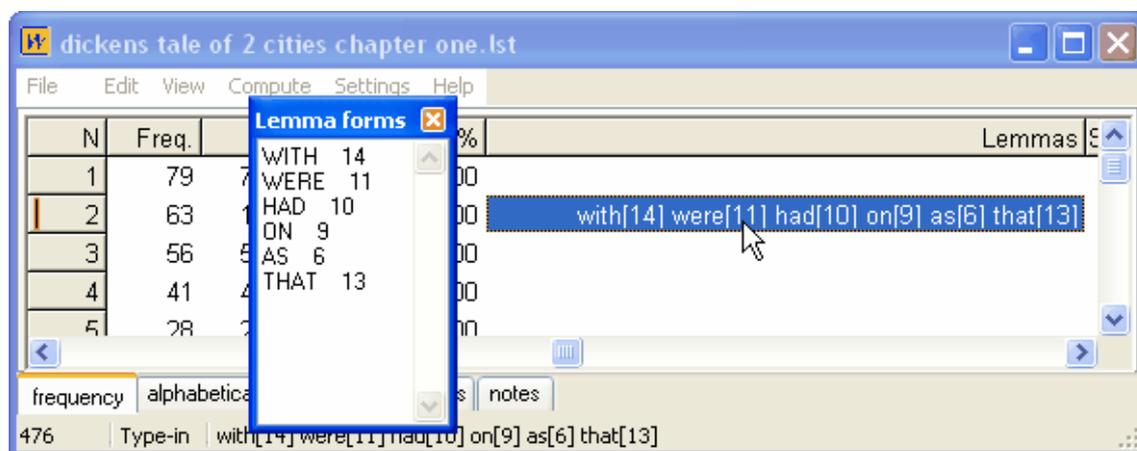
Alternatively you can join up lemmas using a [text file](#) which automates the matching & joining process. The actual processing of the list takes place when you choose the menu option *Match Lemmas* (≡) in WordList, Concord or KeyWords. Every entry in your lemma list will be checked to see whether 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**.

Auto-joining

To speed up this lemmatisation process, you can auto-join any of the entries in your current worldist which meet your criteria.

Can't read all the lemma forms

Double-click on the Lemmas column as in the shot below,



and a window of Lemma Forms will open up, showing the various components.

See also: [Auto-Join](#), [Using a text file to lemmatise](#)

9.10 index lists: uses

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](#)" scores which relate word types to each other.
3. Or you can use an index to see [word clusters](#).

See also [Making an Index List](#), [Viewing Index Lists](#), WordList Help Contents.

9.11 index lists: viewing

In WordList, open an index as you would any other kind of word-list file -- using File | Open. Or, easier in my opinion, in the Controller | Previous lists, choose any index you've made and double-click it. You will see the index as if it were a large word-list.

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 World 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 World.

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.

The screenshot shows a window titled 'index: bnc world.tok' with a menu bar (File, Edit, View, Compute, Settings, Help) and a table of concordance data. The table has columns for N, Word, Freq., %, Texts, %, and emmas. The data is as follows:

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

At the bottom of the window, there are tabs for 'frequency', 'consistency', 'statistics', 'filenames', and 'notes'. The status bar shows '126,550 Type-in'.

You can highlight one or more words or mark them with the  option, then  to get a speedy concordance.

See also [Making an Index List](#), [WordList clusters](#), WordList Help Contents.

9.12 making a WordList Index

To create an index, first use the main [Controller](#) and choose *Adjust Settings | Index*. You will need to specify the file name for the index.

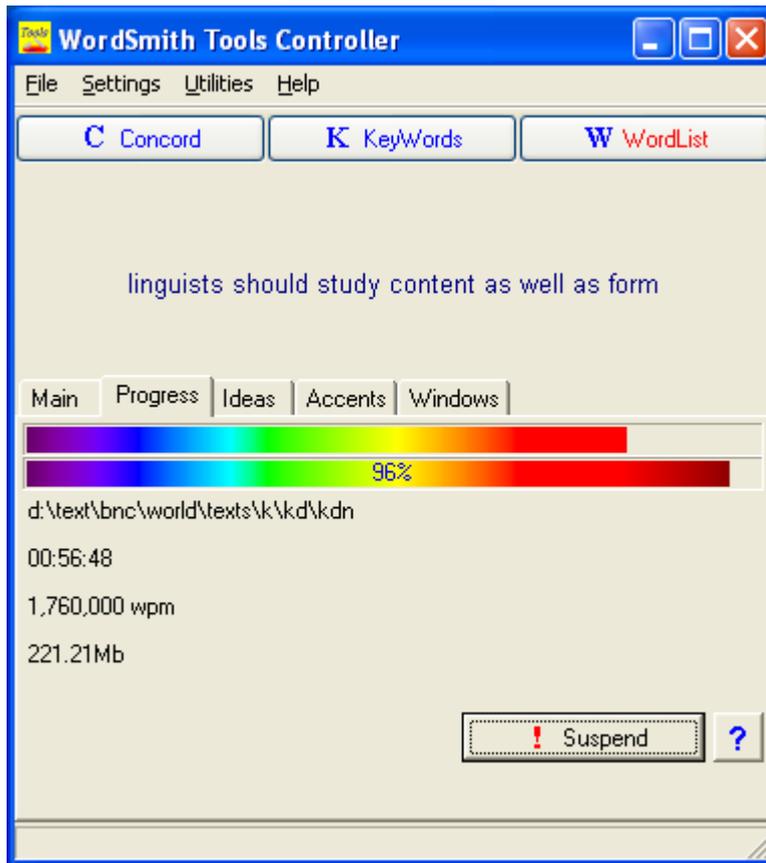
WordList will go through your selected texts and store information about the position of every instance of every word-type.

index files

Two files are created for each index:

- .tok file: a large file containing information about the position of every word token in your text files.
- .typ file: knows the individual word types.

An index permits the computation of [word clusters](#) and [Mutual Information](#) scores for each word type. The screenshot below shows the progress bars for an index of the BNC World corpus; it has taken nearly one hour to do 96% of the work: a rate of about 1.8 million words per minute. The resulting `BNC Words .tok` file was 1.6GB in size and the `BNC Words .typ` file was 26 MB.



adding to an index

To add to an existing index, just choose some more texts and choose File | New | Index. If the existing filename is already in use for an index, you will be asked whether to add more ('Yes') or start it afresh ('No').

See also [Using Index Lists](#), [Viewing Index Lists](#), WordList Help Contents.

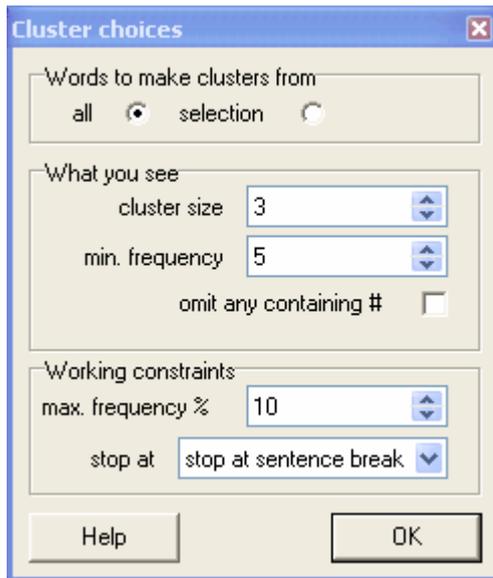
9.13 index clusters

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](#).

How to see clusters...

[Open](#) the index. Now choose *Compute | Clusters*.



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 World), 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 World 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	77.4
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	78.96
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

frequency alphabetical statistics filenames notes
343,617 Type-in LIFE

The process will take time. In the case of BNC World, 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.

Here the user has chosen to see any 3-word clusters that appear 5 or more times.

Working constraints

The "max. frequency %" setting is to speed the process up. 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, 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 World 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](#), offers a number of constraints, such as sentence and other punctuation-marked breaks. 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.

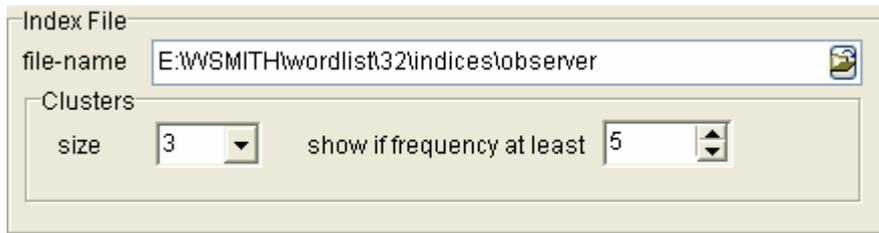
What they look like

The screenshot shows the WordList application window with a menu bar (File, Edit, View, Compute, Settings, Help) and a table of results. The table has columns for N, Word, Freq., %, Texts, and % err. The first row is highlighted in blue.

N	Word	Freq.	%	Texts	% err
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

Below the table, there are navigation arrows and a set of buttons: frequency, alphabetical, statistics, filenames, notes. At the bottom, there is a 'Type-in' field containing 'RABIES'.

Here is a small set of 3-word clusters involving rabies from the BNC World corpus. Some of them are plausible multi-word units. All clusters which appear at least 5 times are shown: to alter that setting, choose *Adjust Settings | Index* in the Controller and set the "show if frequency.." number thus:



See also: [clusters in Concord](#)

9.14 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 wordlist, 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](#) (joining words which belong under one entry, such as *seem/ seems/ seemed/ seeming* etc.)

See also: [searching for an entry by typing](#)

9.15 mutual information scores

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.

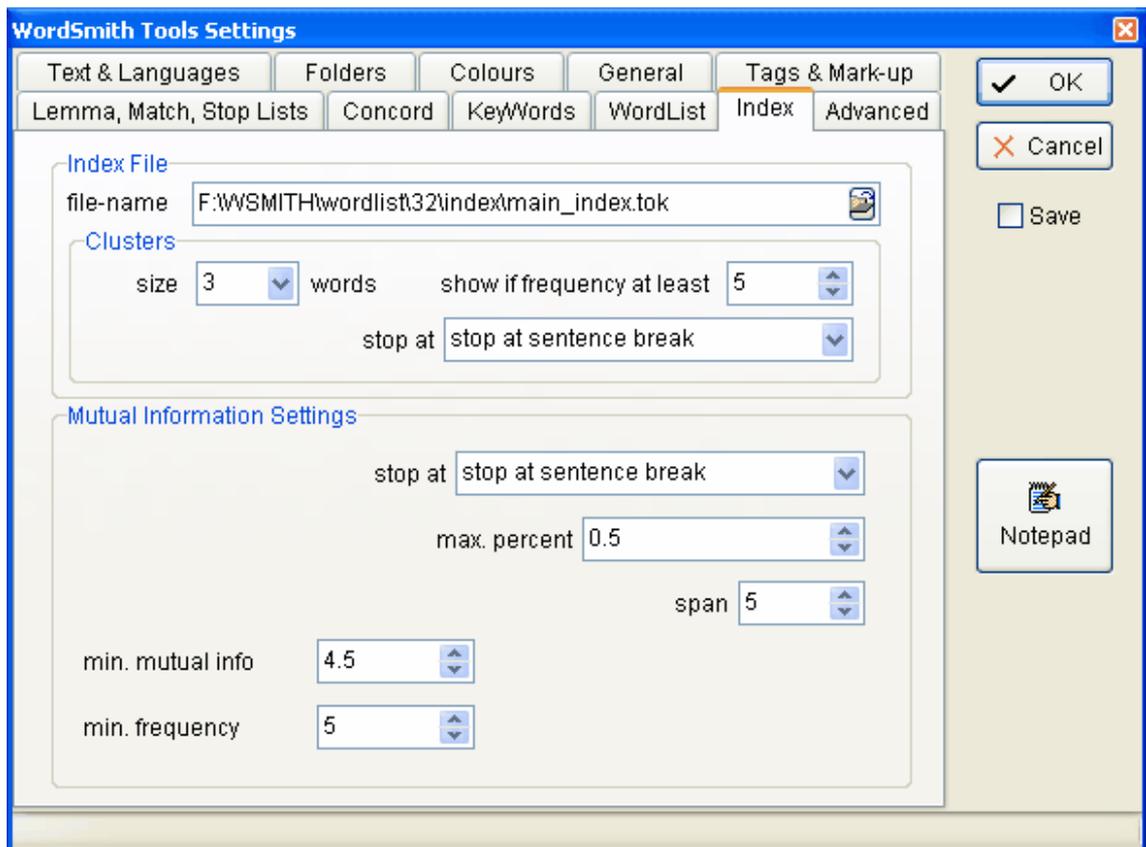
This relationship is bi-lateral: in the case of *kith* and *kin*, it doesn't distinguish between the virtual certainty of finding *kin* near *kith*, and the much lower likelihood of finding *kith* near *kin*.

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](#), 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](#).

Settings

The Mutual Information settings are found in the [Controller](#) under *Adjust Settings | Indexing* or in

a menu option in **WordList**.



stop at: you can choose where you want collocational breaks to be assumed. With the setting above, "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.

max. percent ignores any tokens which are more frequent than the percentage indicated. (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%.)

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*.

min. mutual info. the minimum number which the MI must come up with to be reported. A useful limit is 3.0. Below this, the linkage between node and collocate is likely to be rather tenuous.

min. frequency: the minimum frequency for any item to be considered for the mutual information calculation (default = 5). (If an item occurs only once or twice, the mutual information is unlikely to be informative.)

See also: [Mutual Information Display](#), [Computing Mutual Information](#), [Making an Index List](#), [Viewing Index Lists](#), WordList Help Contents.

See [Oakes](#) for further information about Mutual Information.

9.16 mutual information: computing

In WordList or in Concord

In Concord

MI is not computed by default for a collocate list. To compute MI, you need a word list to supply the relevant data.

Suppose you have made a concordance using all the files in

`c:\wsmith4\text\shakespeare` and have done a concordance on *love*. You get collocates such as *Romeo, hate, the, Juliet, Nurse* etc. All these show a "Relation" (MI) score of "??" because they haven't yet been computed.

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](#)). Make sure the [reference corpus](#) file is what you prefer.

Now choose the menu item  and Concord will use the reference corpus filename. It will look up each of your collocates and compute MI using the information in the reference corpus.

In WordList

To compute Mutual Information (MI) you need a [WordList Index](#). Call up the alphabetical [view](#) of the list.

When you press , you can choose whether to compute MI for selected (highlighted) entries, for all entries, or for those between two initial characters e.g. between A and D.

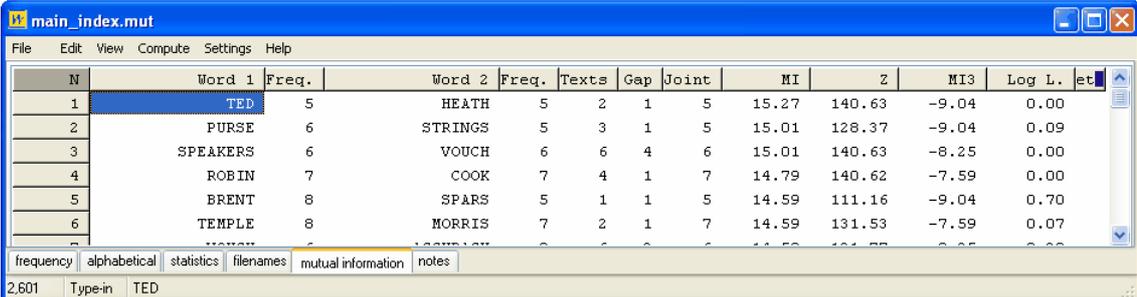
If you wish to select only a few items for MI calculation, you can mark them first (with ).

Computing the MI score for each and every entry in an index takes a long time: at least a couple of hours in the case of an index based on 10 million words. 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! You can always do part of the list (eg. A to D) and later [merge](#) your mutual-information list with another (E to H).

See also [Collocates](#), [Mutual Information Settings](#), [Mutual Information Display](#), [Making an Index List](#), [Viewing Index Lists](#), [WordList Help Contents](#).

9.17 mutual information display

The Mutual Information procedure contains a number of columns:



N	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI	Z	MI3	Log L.	et
1	TED	5	HEATH	5	2	1	5	15.27	140.63	-9.04	0.00	
2	PURSE	6	STRINGS	5	3	1	5	15.01	128.37	-9.04	0.09	
3	SPEAKERS	6	VOUCH	6	6	4	6	15.01	140.63	-8.25	0.00	
4	ROBIN	7	COOK	7	4	1	7	14.79	140.62	-7.59	0.00	
5	BRENT	8	SPARS	5	1	1	5	14.59	111.16	-9.04	0.70	
6	TEMPLE	8	MORRIS	7	2	1	7	14.59	131.53	-7.59	0.07	

Word 1: the word to the left, followed by Freq. (its frequency in the whole index).

Word 2: the word to the right, followed by Freq. (its frequency in the whole index).

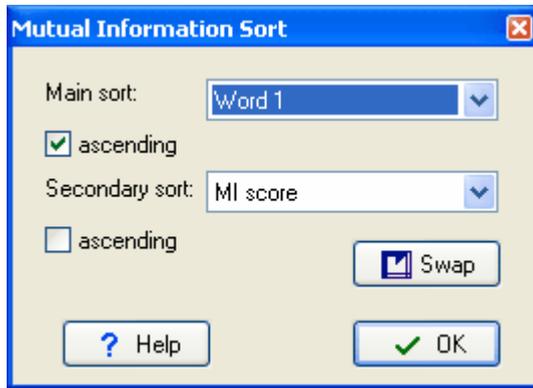
Texts: the number of texts this pair was found in (there were 56 in the whole index).

Gap: the most common distance between Word 1 and Word 2.

Joint: their joint frequency.

In line 2 of this display, PURSE occurs 6 times in the whole index, and STRINGS 5 times. They occur together 5 times -- in other words in this little corpus, **strings** is always part of the phrase **purse strings**. The gap is 1 because **strings** comes 1 word after **purse**. The pair **purse strings** comes in 3 texts.

As usual, the data can be sorted by clicking on the headers. Above, it was sorted by clicking on "MI" first and "Word 1" second.



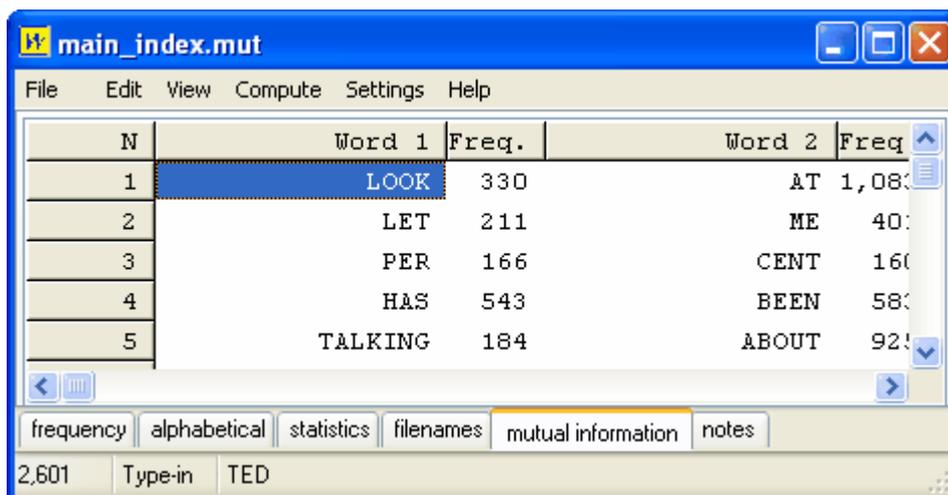
You get a double sort, main and secondary, because sometimes you will want to see how MI or Z score or other sorting affects the whole list and sometimes you will want to keep the words sorted alphabetically and only sort by MI or Z score within each word-type. Press *Swap* to switch the primary & secondary sorts.

Compare this with the display sorted by Z Score ([Oakes p. 163](#)).

N	Word 1	Freq.	Word 2	Freq.
1	TED	5	HEATH	5
2	SPEAKERS	6	VOUCH	6
3	ROBIN	7	COOK	7
4	BRAINTEASER	14	CROCODILE	14
5	RECORDED	55	TRANSMISSION	55

TED HEATH (a UK Prime Minister of the 1970s) is still top and SPEAKERS ... VOUCH still visible, but some other items have moved in.

Here is the display sorted by MI3 Score ([Oakes p. 172](#)):

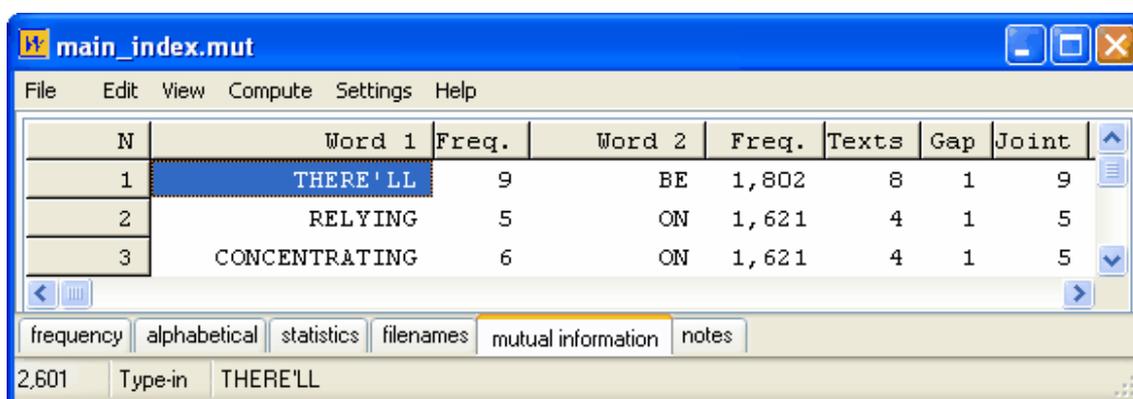


N	Word 1	Freq.	Word 2	Freq.
1	LOOK	330	AT	1,080
2	LET	211	ME	400
3	PER	166	CENT	160
4	HAS	543	BEEN	580
5	TALKING	184	ABOUT	920

2,601 Type-in TED

Much more frequent items have jumped to the top.

Finally, by Log Likelihood ([Dunning, 1993](#)):



N	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint
1	THERE 'LL	9	BE	1,802	8	1	9
2	RELYING	5	ON	1,621	4	1	5
3	CONCENTRATING	6	ON	1,621	4	1	5

2,601 Type-in THERE'LL

Here the Word 2 items are very high frequency ones and we get at colligation (grammatical collocation).

See also: [Mutual Information](#), [Computing Mutual Information](#), [Making an Index List](#), [Viewing Index Lists](#), [WordList Help Contents](#).

See [Oakes](#) for further information about Mutual Information.

9.18 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](#)

9.19 statistics

These include:

- number of files involved in the word-list
- file size (in bytes, i.e. characters)
- running words in the text (*tokens*)
- no. of different words (*types*)
- [type/token ratios](#)
- no. of [sentences](#) in the text
- mean sentence length (in words)
- standard deviation of sentence length (in words)
- no. of [paragraphs](#) in the text
- mean paragraph length (in words)
- standard deviation of paragraph length (in words)
- no. of [headings](#) in the text
- mean heading length (in words)
- no. of [sections](#) in the text
- mean section length (in words)
- standard deviation of heading length (in words)
- the number of 1-letter words

...

the number of **n**-letter words (to see these scroll the list box down)

(14 is the [default](#) 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.

See also : [WordList display](#) (with a screenshot), [Summary Statistics](#), [Starts and Ends of Text Segments](#).

9.20 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](#) (select this before you make the list), and can be in [Unicode](#) or [ASCII](#).

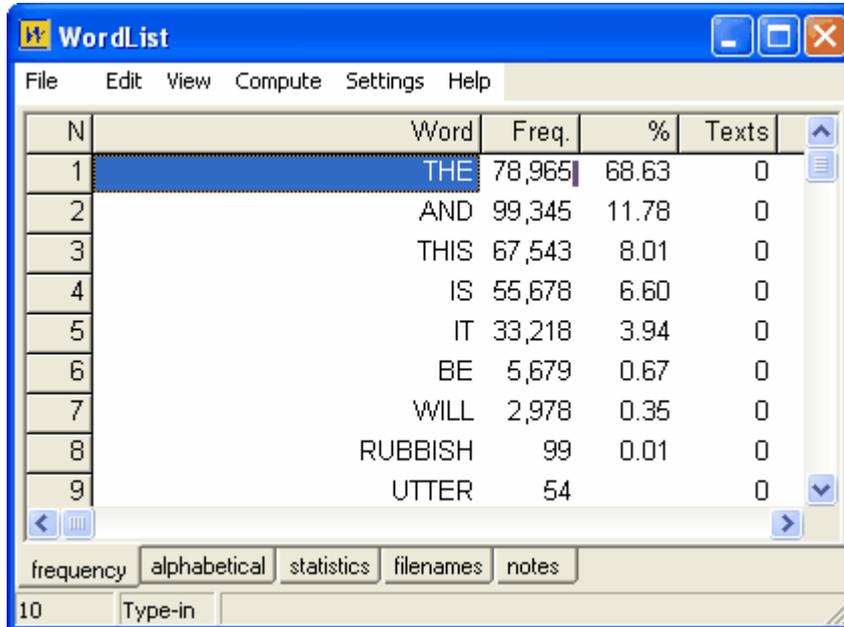
But it must follow a similar format as a [stop list](#) expects, except that following each word there must be a <tab> character and the frequency as a plain number (decimal points will be ignored). Do not use commas as a thousands delimiter as otherwise they'll be interpreted as different words. The words do not need to be in frequency or alphabetical order.

Example

```
; My word list for test purposes.
THIS 67543
IT    33218
WILL  2978
BE    5679
```

COMPLETE 45
 AND 99345
 UTTER 54
 RUBBISH 99
 THE 578965
 IS 55678

You should get results like these.



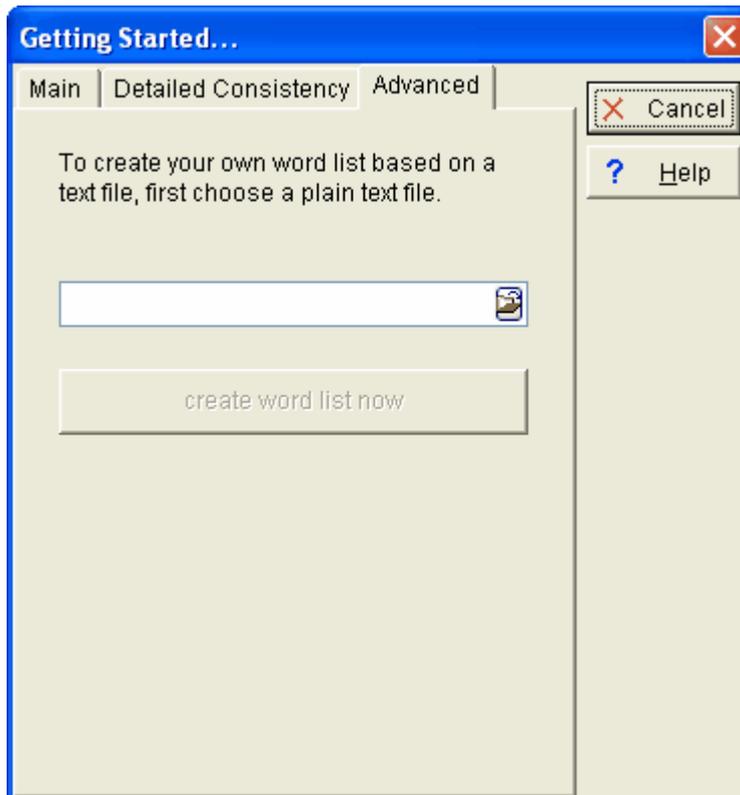
The screenshot shows the WordList application window with a menu bar (File, Edit, View, Compute, Settings, Help) and a table of word statistics. The table has columns for Rank (N), Word, Frequency (Freq.), Percentage (%), and Texts. The 'frequency' tab is selected at the bottom.

N	Word	Freq.	%	Texts
1	THE	78,965	68.63	0
2	AND	99,345	11.78	0
3	THIS	67,543	8.01	0
4	IS	55,678	6.60	0
5	IT	33,218	3.94	0
6	BE	5,679	0.67	0
7	WILL	2,978	0.35	0
8	RUBBISH	99	0.01	0
9	UTTER	54		0

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.

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,



one for [Detailed Consistency](#), and another (*Advanced*) for creating a word list using a plain text file.

Choose your .txt file and press *create word list now*.

9.21 type/token ratios

If a text is 1,000 words long, it is said to have 1,000 "tokens". But a lot of these words will be repeated, and there may be only say 400 different words in the text. "Types", 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 uses a different strategy for computing this, therefore. The standardised type/token ratio (STTR) is computed every *n* words as Wordlist goes through each text file. By [default](#), *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](#) to any number between 100 and 20,000.

What STTR actually counts

Note: The ratio is computed a) counting every [different form](#) as a word (so **say** and **says** are two types) b) using only the words which are not in a [stop-list](#) c) those which are within the length you have specified, d) taking your preferences about [numbers](#) and [hyphens](#) 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](#) (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!)

9.22 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 (*Adjust Settings | WordList | Case Sensitivity* in the [Controller](#)).

When you first see your case-sensitive list, it is likely to appear all in UPPER CASE. Press *Ctrl/L* or choose the [Layout](#) menu option () to change this.

9.23 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 wordlist 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](#).

See also: [Text Characteristics](#), [Stop Lists](#), [Setting Defaults](#)

9.24 sort order

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.

To choose one of the special sorts specified below, press F6 or Ctrl/F6 or Shift/Ctrl/F6. Or choose the appropriate menu option.

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](#) 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**.

Reverse Word Sort

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

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](#).

See also: [Concord sort](#), [KeyWords sort](#), [Editing entries](#); [Accented characters](#); [Choosing Language](#)

9.25 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	1,270	5.52	7	100.00		
2	TO	775	3.37	7	100.00		
3	THAT	654	2.84	7	100.00		
4	AND	523	2.27	7	100.00		
5	OF	495	2.15	7	100.00		
6	YOU	447	1.94	7	100.00		
7		417	1.81	7	100.00		

Here you see the top 6 words in a word list based on 7 interviews. There are 2,479 words altogether but in the screenshot we can only see the first few. The Freq. column shows how often each word cropped up (**THE** appeared 1,270 times in the 7 texts), and the % column tells us that 1,270 represents 5.52% of the running words in the 7 texts. The Texts column shows that **THE** comes in 7 texts, that is 100% of the texts used for the word list.

N	Word	Freq.	%	Texts	%	emmas	Set
4	ABERDEEN	1	0.00	1	14.29		
5	ABILITY	1	0.00	1	14.29		
6	ABLE	18	0.08	5	71.43		
7	ABOLISHING	1	0.00	1	14.29		
8	ABOUT	70	0.30	7	100.00		
9	ABOVE	1	0.00	1	14.29		
10	ABOLUTE	1	0.00	1	14.29		

The Alphabetical listing also shows us some of the words but now they're in alphabetical order.

ABLE comes 18 times altogether, and in 5 of the 7 texts. **ABOUT**, on the other hand, comes in all 7 texts.

Now let's examine the statistics.

	N	0	1	2	3
text file	overall	31.5.98.txt	alex salmond	15.3.98.txt	4.6.98.tx
file size	151,474	23,489		20,216	37,530
tokens (running words) in text	22,992	3,485		3,089	5,730
tokens used for word list	22,992	3,485		3,089	5,730
types (distinct words)	2,479	824		674	924
type/token ratio (TTR)	10.78	23.64		21.82	16.10
standardised TTR	37.15	36.37		37.37	35.00
standardised TTR std.dev.	58.95	49.00		46.95	53.50
standardised TTR basis	1,000	1,000		1,000	1,000
mean word length (in characters)	4.49	4.52		4.55	4.44
word length std.dev.	2.42	2.48		2.58	2.30
sentences	928	147		132	230
mean (in words)	24.78	23.71		23.40	24.20
std dev	19.73	20.38		15.28	18.70

frequency alphabetical **statistics** filenames notes

2,479 Type-in

In all 7 texts, there are 2,749 word types (as pointed out above). The total running words is 22,992. Each word is about 4.49 characters in length. There are 928 sentences altogether, on average 24.78 words in length. In the text of the interview with Alex Salmond, there are only 674 different word types and that interview is only just over 3,000 words in length. This is explained in more detail in the [Statistics](#) 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
22	TRAFFIC	3	0.01	2
23	SPECIFIC	4	0.02	4
24	TERRIFIC	1	0.00	1
25	MAGIC	1	0.00	1
26	STRATEGIC	2	0.01	2
27	PUBLIC	31	0.13	4
28	DYNAMIC	1	0.00	1

frequency reverse alphabetical statistics filenames notes

2,479 Type-in TRAFFIC

To do a reverse alphabetical sort, I had the Alphabetical window visible, then chose *Edit | Reverse Word sort* in the menu. To revert to an ordinary alphabetical sort, press F6.

See also : [Consistency](#), [Lemmatisation](#)

9.26 WordSmith controller: WordList settings

These are found in the main [Controller](#) under *Adjust Settings | WordList*.

This is because some of the choices -- e.g. [Minimum & Maximum Settings](#) -- may affect other Tools.

There are 2 sets : What you Get and What you See.

WHAT YOU GET

Word Length & Frequencies

See [Minimum & Maximum Settings](#).

Standardised Type/Token

See [WordList Type/Token Information](#).

Tags

If you want to include tags in a word list, you need to set up a [Tag File](#) first. Then ensure that "Include Tags" is checked, so that you will see the tags in the word list. Tags are not counted in any statistics based on a running word count or number of tokens or types. What you will see for each is its frequency, that frequency as a percentage of the running words excluding tags, and the number of texts it is in.

WHAT YOU SEE

Case Sensitivity

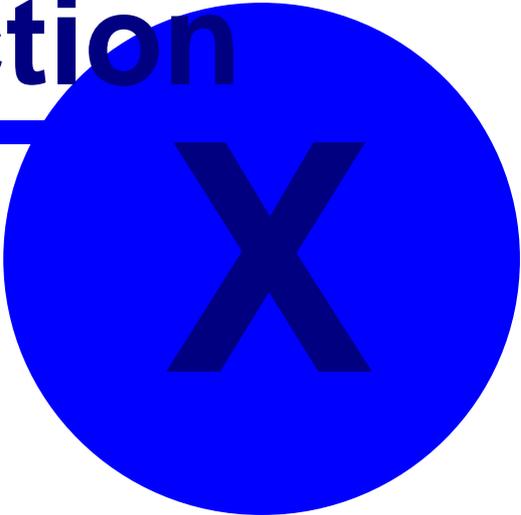
Normally, you'll make a case-insensitive word list. If you wish to make a word list which distinguishes between the, The and THE, activate [case sensitivity](#).

See also: [Using Index Lists](#), [Viewing Index Lists](#), WordList Help Contents, [Computing word list clusters](#).

WordSmith Tools

Utility Programs

Section



X

10 Utility Programs

10.1 Convert Data from Previous Versions

10.1.1 Convert Data from Previous Versions

As Oxford 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\(s\)](#) were selected for a concordance, [notes](#) 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 to the new format for the current version.

16

32

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.

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 and press Go or <Enter>.

How it works

WebGetter visits the Search Engine specified in the second box and downloads the first 100 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](#)). 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](#), [Display](#), [Limitations](#)

10.2.2 settings

These are

- where the texts are to be stored. The folder you specify will act as a *root*. That is, if you specify `c:\temp` and search for "besteiorl", results will be stored in `c:\temp\besteiorl`. If

you do another search on say "Oxford WordSmith Tools", results for that will go into
 c:\temp\WordSmithTools.

- timeout: the number of seconds after which WebGetter robot stops trying a given webpage if there's no response. Suggested value: 20 seconds.
- max simultaneous: webgetter works by sending robots out simultaneously, each one requesting a different web page. Suggested value: 20. That is, up to 20 are being downloaded at once.
- language: you specify the language you require.
- 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.

Search Engines

Download a choice of search engines by pressing Engines. This gets the latest information about each search engine from www.lexically.net/downloads/searchengines.htm.

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.

The grid of settings

This contains:

name	The Name to appear above, in the list of Search Engines
ignore	Websites not to visit when downloading (as opposed to requesting a list). That is, when WebGetter gets a page from Google, it only wants Google's list, not more Google web-pages.

URL	The URL where the Search Engine is found.
Searchstring	The search word syntax
Max	How many hits to try for on each contact
Next	
Language	Required language
Other	

The search word is specified more or less just as you do when you use the same Search Engine yourself. Few advanced settings for each Search Engine are used; you can try your own preferences by typing in the grid, in the Searchstring column. Learn each Search Engine's current settings by simply trying it and then adapt the Searchstring accordingly. Some Search Engines want to set cookies on your PC and this might cause a failure to download.

You can see the address line in the Advanced tab; WebGetter attempts to tell the Search Engine the search-word, the maximum number of hits to show per contact, what language to use, and how to get more.

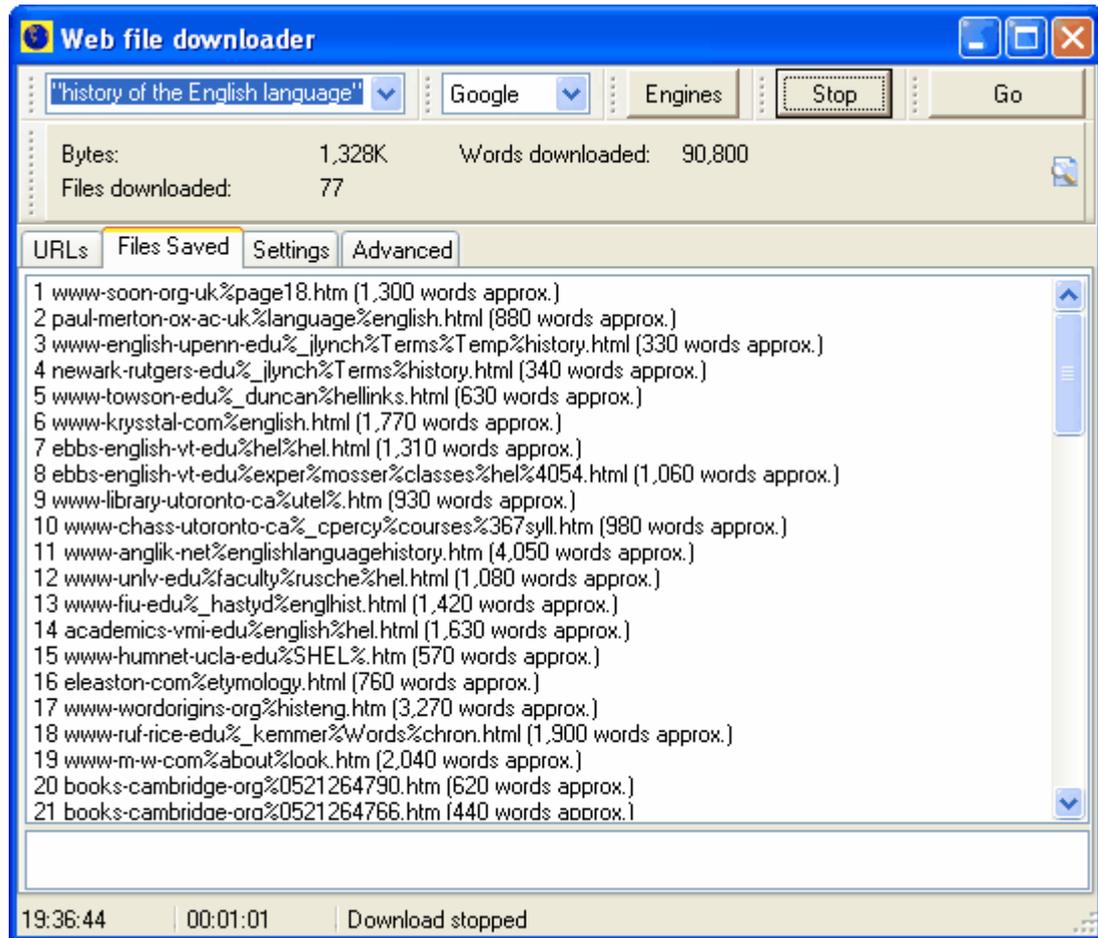
See also: [Display](#), [Limitations](#)

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. If dark blue, they were saved to disk. Above, you will see the bytes visited, and every time a file which meets your requirements is stored, you'll see the number of files and number of words go up. At the bottom, the current time and elapsed time.

There is a tab giving access to a list of the successfully downloaded files.

Here is a partial list of what I got with a broadband connection, in 1 minute & 1 second, with the search term "history of the English language" (*with* quotes).



As you can see, about 1.3MB of web-pages were examined, and 90,000 words (1.1MB) were found worth saving, with the default settings (they each had to be at least 10K in size and have 300 words). In that time I got a couple of time-outs, presumably because 20 seconds isn't long enough for some websites or servers which are slow and ponderous.

See also: [Settings](#), [Limitations](#)

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](#) to be demanding about what you download.

See also: [Display](#)

10.3 Languages Chooser

10.3.1 Overview

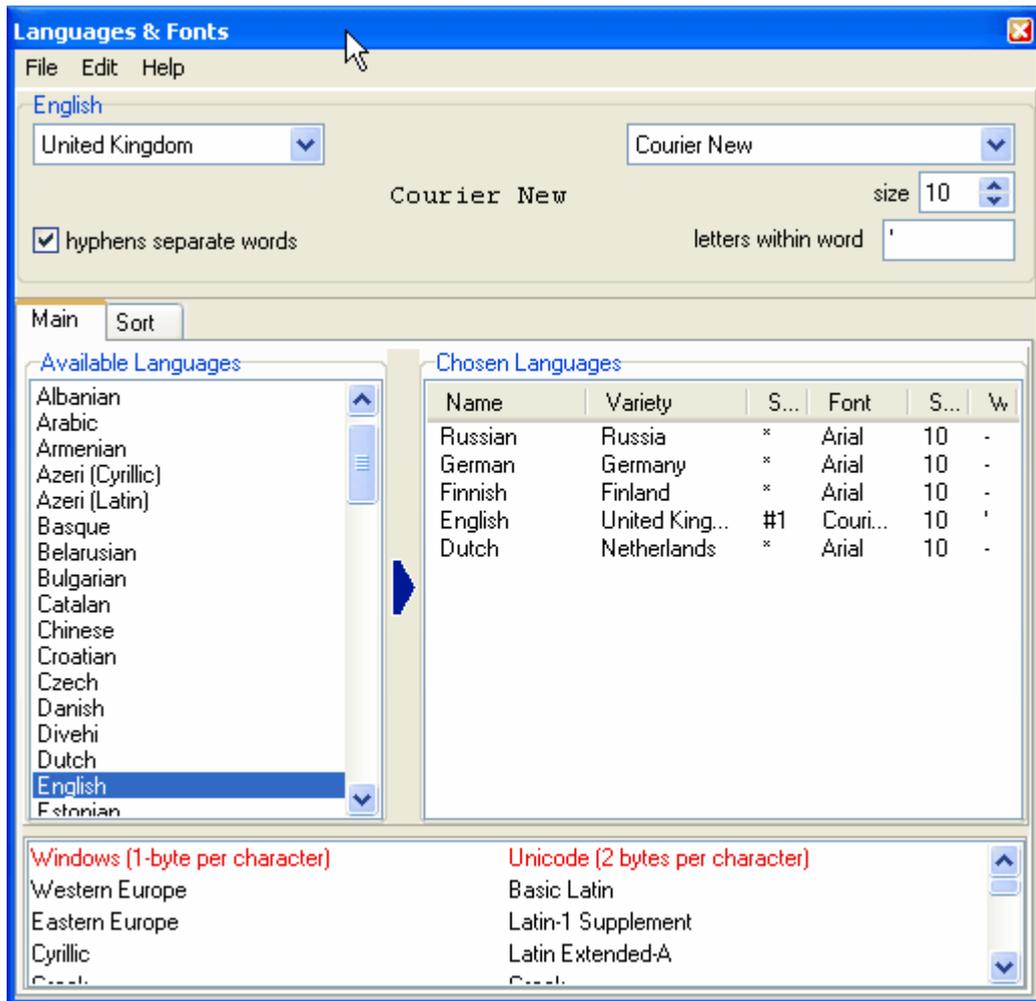


A tool for selecting Languages which you want to process.
You will probably only need to do this once, when you first use Oxford WordSmith Tools.

How to get here

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

What you will see may look like this:



5 languages have been chosen already.

At the bottom you will see what the current font can handle, in terms of Windows ANSI or Unicode text. The Courier New font on the PC this was done on can handle characters in Windows for Western and Eastern Europe, Cyrillic etc., as well as several ranges within the Unicode standard.

See also : [Language](#), [Font](#), [Sort Order](#), [Other Languages](#), [saving your choices](#)

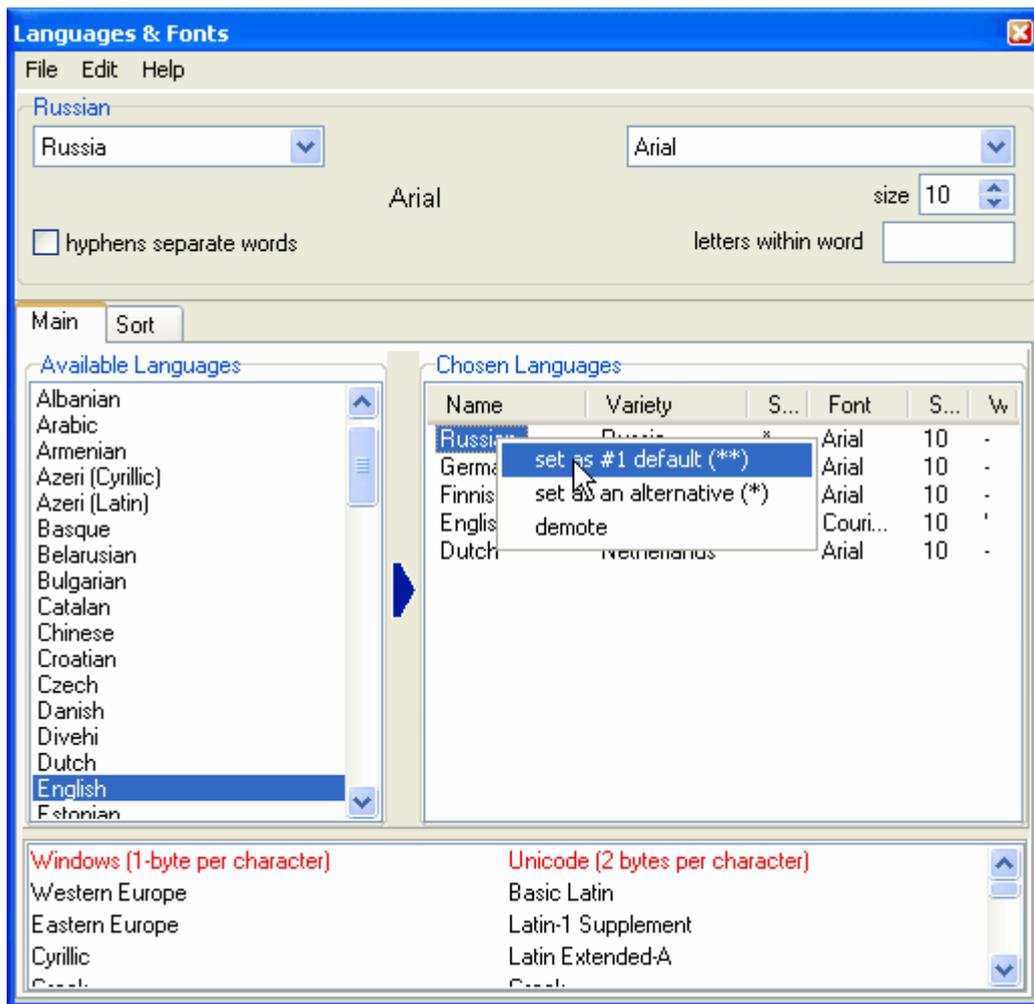
10.3.2 Language

How to get here

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

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.)



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 *. For each Chosen Language, you can specify any symbols which can be included within a word, e.g. the apostrophe in English, 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).

To change the status of a chosen language, right-click. This user is about to make Russian 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.

Each time you change language, the list of [fonts](#) available changes, and the [sorted words](#) will change their appearance. The window at the bottom shows which characters can be supported in Unicode or 1-byte format by the highlighted language.

Some languages do not mark [word-separators](#).

See also : [Other Languages, saving your choices](#)

10.3.3 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.

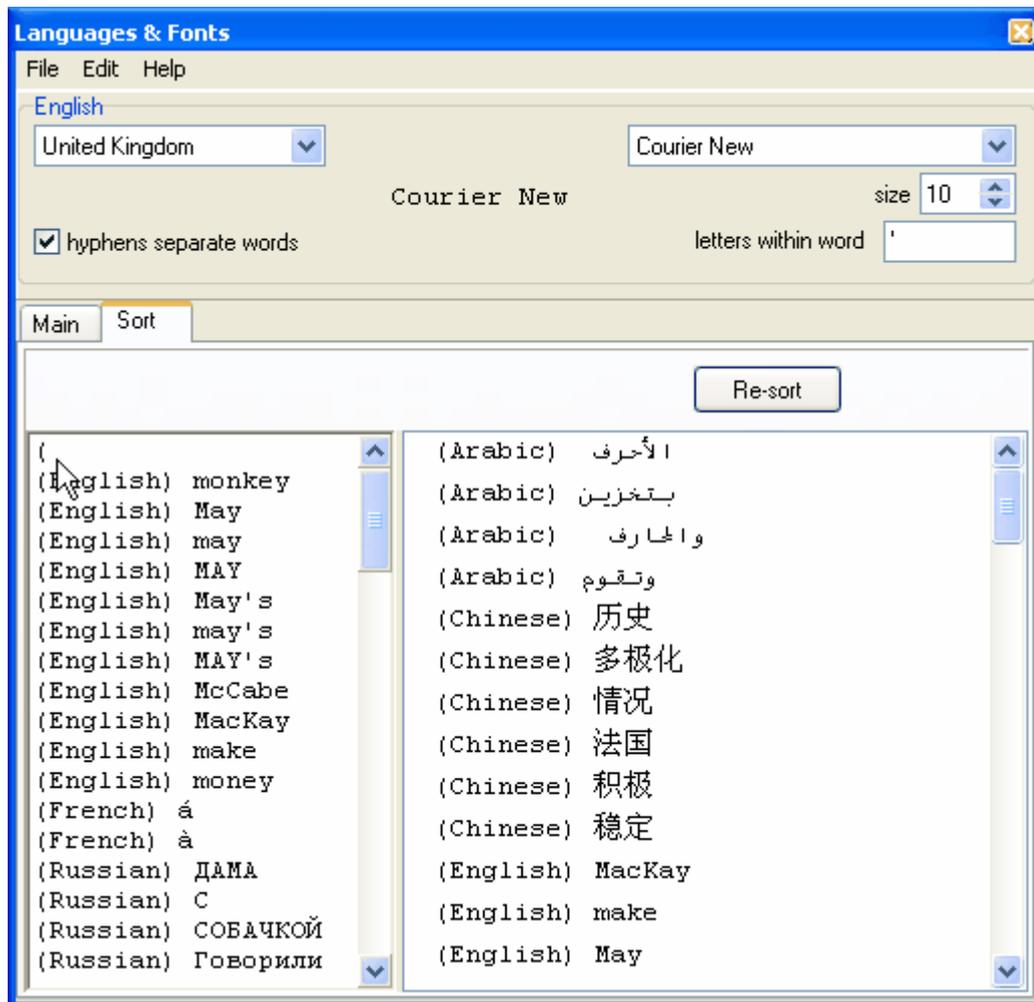
Click here for more on [Unicode](#).

See also : [Language](#), [Sort Order](#), [Other Languages](#), [saving your choices](#)

10.3.4 Sort Order

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

The display



- You will see 2 windows below "Resort" -- the one at the left contains some words in various languages; you can add your own. The cursor in the screenshot shows where a user is about to type, having already typed "(" . 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](#), [Font](#), [Other Languages](#), [saving your choices](#)

10.3.5 Other Languages

To work on a language not in the list, press *Edit* and base your new language name 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](#), [Font](#), [Sort Order](#), [saving your choices](#)

10.3.6 saving your choices

Save your results before quitting, so that next time Oxford 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 `\wsmith4\language_choices.ini`.

See also : [Language](#), [Font](#), [Sort Order](#), [Other Languages](#)

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](#), [choosing your files](#), [output](#), [rules and settings](#), [running the program](#).

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          75,432
WAS          9,895
```

or

1	THE	75,432
2	WAS	9,895

You can make such a list using [WordList](#). For example, select (highlight) the columns containing the word and its frequency, press the ".txt" button, then

- Clear the "Number each line" box
- Rows to save = "all" (but if it shows 0-xxx change 0 to 1)
- Columns to save = "any highlighted"

See also : [aim](#), [choosing your files](#), [output](#), [rules and settings](#), [running the program](#).

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.
- If it has numbered lines, check the ".txt is pre-numbered" box.
- If it has a header (WS3 will by default produce 3 lines of header information) make sure you have set the "Header lines to skip" box to the right number.
- You must specify where to save your results. The results will show all the typos and minimal pairs which the program finds.
- Choose also,
 - whether to number the list of results
 - whether to show the frequencies of possible typos
 - whether to show the rule which generated the result.

See also : [aim](#), [requirements](#), [output](#), [rules and settings](#), [running the program](#).

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]
```

See also : [aim](#), [requirements](#), [choosing your files](#), [rules and settings](#), [running the program](#).

10.4.5 rules and settings

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 start-of-word match

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!

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.

all 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)

Save Current Settings (saves your choices of file and rules)

Get Saved Settings (restores your last-saved choices)

See also : [aim](#), [requirements](#), [choosing your files](#), [output](#), [running the program](#).

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.

If you want to stop in the middle, press "Stop".

The status bar at the bottom of the screen shows how many words have been found in the word-list, the time elapsed, and time estimated to completion of the whole task.

You can press "Results" to see your results file, when you have finished.

Finally, "Quit".

See also : [aim](#), [requirements](#), [choosing your files](#), [output](#), [rules and settings](#)

10.5 File Utilities

10.5.1 index



This sub-program supplies a few file utilities for general use:

[Compare Two Files](#)

[File Chunker](#)

[Find Duplicates](#)

[Rename](#)

Find Holes: for "[holes](#)" in text files

[Splitter](#)

[Joiner](#)

10.5.2 Splitter

10.5.2.1 aim of Splitter

This is a sub-program for splitting large files into lots of small ones. **Splitter** needs to know:

End of Text Separator

The symbol which will act as an end-of-text separator: eg. **[FF]** or **<end of story>** or **</Text>** or **!#** or **[FF*]** or **[FF?????]**

Restrictions:

- 1 The 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 end-of-text separator may not be a [wildcard](#) such as **#**, ***** or **?**.
- 4 ***** and **#** may occur only once each in the end-of-text separator.

Splitter will create a new file every time it encounters the end-of-text marker you've specified.

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 = 2 and 30,000). Only files within these limits will be saved. This feature is useful for extracting files from very large CD-ROM files. The default of 2 is to avoid getting little text files e.g. from newspaper *News in Brief* stories, but if you do want small texts, then set this to 1.

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. This is because often the first line after your end-of-text symbol will contain some kind of header. If

you don't want it to insert < and > around the line, leave the checkbox un-checked.

Title Line

If you know which line of your texts always contains the title for the sub-textin question, set this counter to that number, otherwise leave it at 0.

See also: [Joiner](#), [Filenames](#), [Wildcards](#), [The buttons](#), Text Converter index.

10.5.2.2 Splitter: filenames

Splitter will create lots of small files based on your large one(s).

It creates new [filenames](#) on the following basis:

A folder based on the name of the source file is created. 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.

Thus a large file called **HELLO.DAT** will split up into a number of small ones:

```
\HELLODAT\1.txt
\HELLODAT\2.txt
...
\HELLODAT\1\512.txt
```

etc.

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](#).
3. CD-ROM files when copied to your hard disk may be read-only. You can change this attribute using Text Converter.

10.5.2.3 Splitter: wildcards

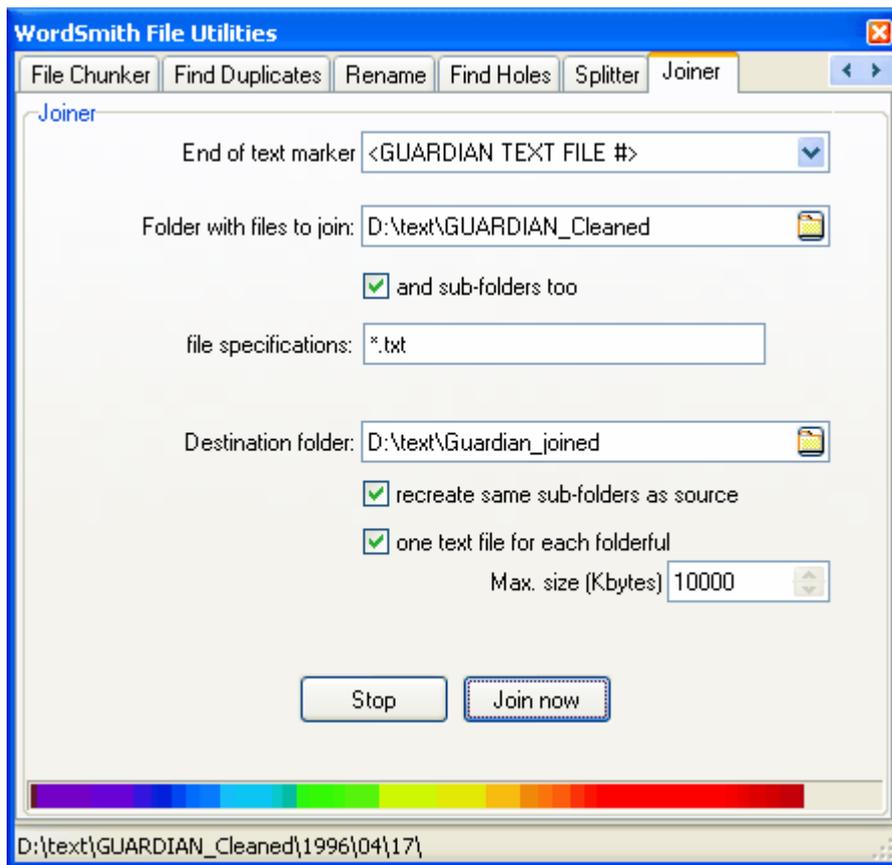
- # The hash symbol, #, is used as a wildcard to represent any **number**, so **[FF#]** would find **[FF3]** or **[FF9987]** but not **[FF]** or **[FF 9]** (because there's a space in it) or **[FFhello]**.
 - * The asterisk represents any **string**, so **[FF*]** would find all of the above. * is used 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](#)

10.5.3 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

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 is 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 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, Text Converter index.

10.5.4 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](#) 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](#), [find duplicates](#), [rename](#)

10.5.5 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 to floppy disks or cdroms. Otherwise how can you get a 5MB file onto 3 or 4 floppy disks and transfer it to another pc?

Naturally on the other pc, you will 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. If to A: you will be asked to put in a new formatted floppy for each chunk.)
3. Specify the "size of each chunk" (default = 1,400K, which fits on a floppy)
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](#), [find duplicates](#), [rename](#)

10.5.6 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.

This program only checks whether the files it is comparing have the same name. (You could use [Compare 2 Files](#) to see whether they are in fact identical.) 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.

See also : [compare two files](#), [file chunker](#), [rename](#)

10.5.7 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.

How to do it

Specify your Folder, whether sub-folders will also be processed, and the kinds of file you want to handle.

The default is *All files *.**.

Also specify a "mask for new name" and a starting number.

For example, with a mask of SUN and start number 0, the first file found, let's say it was originally `Quite_a_long_and_Complicated_file.txt` will be renamed `SUN0.txt`.

The next file would be `SUN1.txt`, and so on. (If the next was

`Quite_a_long_and_Complicated_file.htm`, that would become `SUN2.htm`).

When you press "Find Files", you will see a list of all files meeting these choices. If you now press "Rename" each one will be renamed according to your settings.

See also : [compare two files](#), [file chunker](#), [find duplicates](#)

10.6 Text Converter

10.6.1 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](#) so that **Text Converter** knows which symbols or strings to convert. It operates under Windows and saves using the Windows [character set](#), but will convert text using DOS or Windows character sets. You can use it to make your text files suitable for use with an Internet browser such as Netscape.

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](#) 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](#).

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](#).

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.

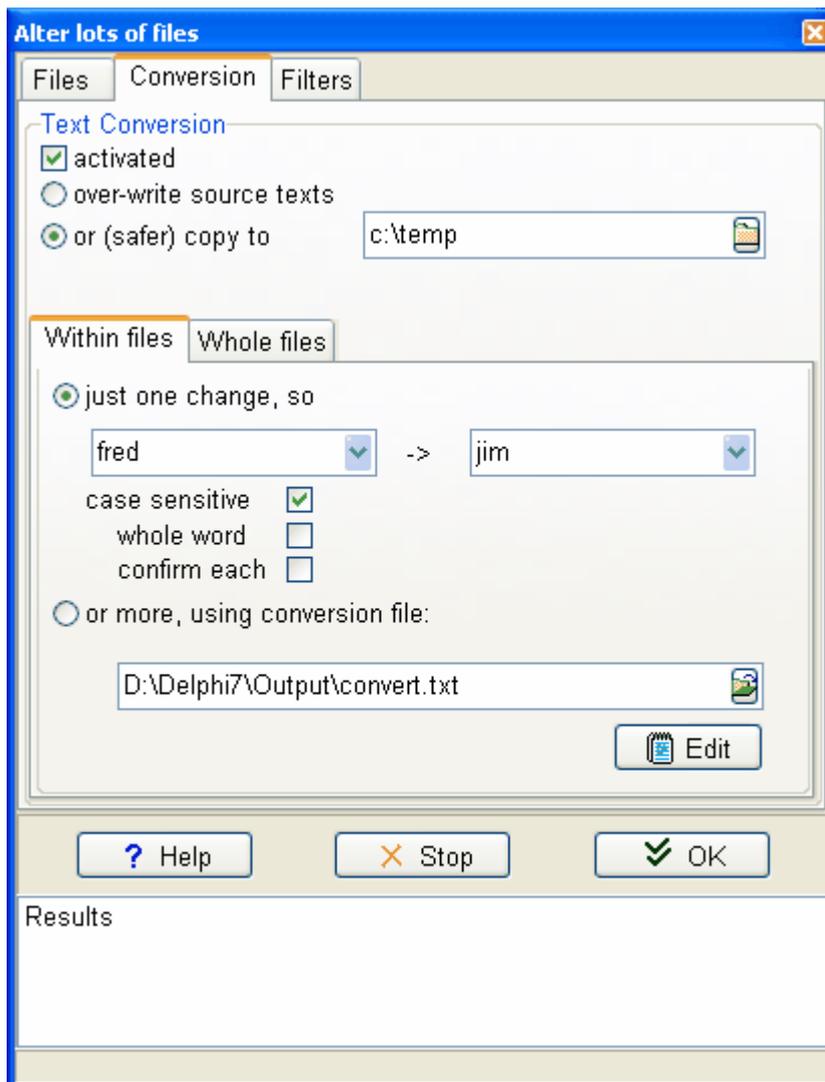
See also: Text Converter Contents, [The buttons](#)

10.6.2 Text Converter: settings

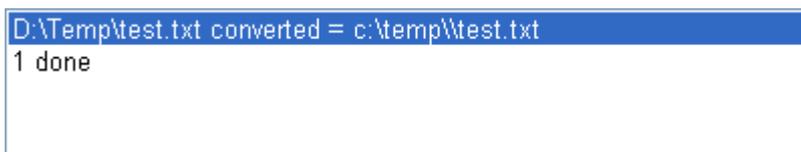
1. Choose *Files* (the top left tab). Decide whether you want the program to process sub-folders of the one you choose. There is no limit to the number of files Text Converter can process in one operation.
2. Click on the *Conversion* tab, and:
3. Decide whether you want to make copies of the text files, 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.
4. 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](#).
5. Or in the Whole Files section you can choose simply to update legacy files by choosing
Dos to Windows,
Unix to Windows,
MS Word .doc to .txt or
into Unicode.
6. 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](#) tab.

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!

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.



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 Cancel button or press Escape.



Right-click to see the source or the converted result file.

See also: [Text Converter Contents](#).

10.6.3 Text Converter: syntax

The syntax for a [Conversion File](#) 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.

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, {CHR(12)} is a printer form-feed. 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.

Wildcards (*,?,# and ~)

* You can use the asterisk as a wildcard. Thus "<*>" -> "" will delete any string in < > brackets from your text. "<head */head>" will delete any string starting "<head " and ending "/head>", even if there are hundreds of characters between them. The default search distance is 1,000 characters, with a maximum of 25,000. (The text is read chunk by chunk into a 30,000 character buffer, so the maximum will work fine at the start of the text; after this only 1,000 characters of search-space are guaranteed.) As deleting a lot of text can get rid of more text than you expect if the text is not properly marked up in the first place, you will probably need to over-ride the default search distance by specifying it in brackets, e.g. "<head*(100)/head>". The asterisk may not be the first or last symbol between the double quotation marks in the search-string.

The asterisk also retains up to 1,000 characters. "<div*(100)>" remembers all the characters up to > and can use them in the replacement: Thus "<div*(100)>" -> "[section *]" will produce [section 1 They Meet Again] if the original has <div1 They Meet Again>. "<div*>" will do the same thing but would allow up to 1,000 characters' search for the >.

Use # to symbolise any number. "<div#>" will find <div1>, <div2> , <div468>, etc. If # is in the replacement too, the exact same number will be used in the replacement. Thus "<div#>" -> "[section #]" will produce [section 468] if the original has <div468>.

? The question mark stands for any single character, except a space. Up to ten ?s can be used in the replacement string to reproduce the character referred to by the ?s in the search-string.

~ The tilde means except. ~"<p>" "<*>" -> "" means delete everything in between angle brackets, except a case of <p>.

Use {CHR(42)} if you need to refer to *, {CHR(35)} for #, {CHR(63)} for ? and {CHR(126)} for ~.

Whole word, case Insensitive, Confirm, redundant Spaces

/C stops to confirm you wish to go ahead before each change.

/W does a whole word search (ensuring the alteration only happens if there's a [word separator](#) on either side) (/W "the" finds the but not other or then or 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.

/IWC "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.

Additions (/A, /T and {v})

/A means add text. /A "Ulan" START inserts Ulan at the start, /A "Bator" END inserts

Bator at the end of the text. See `\wsmith4\convert.txt` to see one in use.

`/T` means add title. So `/T "<title>*</title>" -> "***` looks for `<title> ... </title>` and if it's found, inserts the wording given into the file. This will make your browser show the title at the top of the screen.

`{v="}` means remember this and use it in another line of the conversion file when you find `{v}`.
`"26 Dec." -> "Boxing Day" {v="Xmas"}` stores the reference `Xmas` and `"1 May" -> "Mayday" {v="after Easter"}` stores `after Easter` for use in a later line, such as `"/celebration/" -> "{v}"`. Assuming that your text has a mention of `26 Dec.` and `1 May`, this example, on finding `/celebration/` in the text, will put `Xmas` if the most recent mention in the text was `26 Dec.` and `after Easter` if the most recent mention was `1 May`.

See `\wsmith4\convert.txt` to see examples in use.

See also: Text Converter Contents.

10.6.4 Text Converter: 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 *Expect*.

How to do it

1. Click on the *Filters* tab (at the top).
2. Now the *Activated* checkbox.
3. Specify a word or phrase the text must contain. This is case sensitive.
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
 - copy to a certain folder, or
 - move to that folder, or
 - delete the file (careful!).

You can also decide to build a sub-folder based on the word or phrase you chose in #3. And you may have the program add `.txt` (useful if as with the BNC there are no file extensions).

See also: Text Converter Contents.

10.6.5 Text Converter conversion file

Prepare your Text Converter conversion file using a [plain text](#) editor such as Notepad. You could use `\wsmith4\contvert.txt` as a basis.

If you have [accented characters](#) 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](#), [syntax](#), Text Converter Contents.

10.6.6 Text Converter: sample conversion file

You could copy all or part of this to the [clipboard](#) and paste it into notepad.

```
[ comment line -- put whatever you like here, it'll be ignored ]

[ first a spelling correction ]
"responsible" -> "responsible"

[ now let's change brackets from < > to [ ] and { } to ( ) ]
"<" -> "["
">" -> "]"
"}" -> ")"
"{" -> "("
/S
[ that will clear all redundant spaces]
```

The file `\wsmith4\convert.txt` is a sample conversion file for use with British National Corpus text files.

See also: Text Converter Contents.

WordSmith Tools

Viewer and Aligner

Section



XI

11 Viewer and Aligner

11.1 purpose

This is a program for showing your text or other files, highlighting words of interest. By default you will see them in [plain text](#) format, but you can view [SGML- or HTML-](#)formatted text in a variety of ways, or choose to see any accented characters as numerical codes instead, e.g. so as to examine the codes for [accented characters](#). There are a number of [settings](#) and [options](#) you can change.

Viewer can be used to skim through your texts at a speed which you can set.

Or you can use it simply to [edit](#) your text file with any symbols such as £ converted to the Windows character set, and optionally numbering sentences or paragraphs, by [saving as a text file](#).

You can also use it to produce an [aligned](#) version of 2 texts, with alternate sentences or paragraphs from each of them.

Viewer is a separate utility, but linked to the other Tools: it will be called up whenever you need to return to the source text, e.g. to see your [key words](#) in context.

See also: [Viewer settings](#), [Viewer options](#)

11.2 aligning with Viewer

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](#) 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: [The buttons](#)

11.3 aligning and moving

You will see one text in one colour and the other in another. (Colours can be changed in the *Settings* menu.)

You may well want to alter sentence ordering. The translator may have used three sentences where the original had only one.

adjusting by dragging with the mouse

To alter sentence order, just drag a sentence to the next one above of the same colour.
Finally you will want to [save \(F2\) the results](#).
See also: Viewer contents

11.4 editing

While **Viewer** is not a full word-processor, some editing facilities have been built in to help deal with common formatting problems:

Edit (

Trim extra spaces: 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.

Find lower-case lines: this identifies cases 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.)

[Find short lines](#)

[Insert tags](#)

You will then want to save (F2) your text.

You can also:

- open a new file for viewing (you can open up to 10 text files within **Viewer**)

- copy a text file to the [clipboard](#) (select, then press Control+Ins)

- print the whole or part of the currently active text file

- change the current [character set](#)

- search for words or phrases (press F12)

- skim** (this flips through the text a screenful at a time. The cursor changes shape to show that you're in skim mode. You can alter the speed with a small horizontal scrollbar and stop skimming at any time by clicking on the text o pressing any key.)

11.5 numbering sentences & paragraphs

You can use the **Viewer** 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](#).

To choose between sentence or paragraph mode, see [Settings](#).

See also: Viewer contents

11.6 options

Mode: Sentence/Paragraph

This switches between Sentence mode (the default) 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](#)) work in the same way in either mode. If you re-save the text, you will probably prefer paragraph mode; this will put an <Enter> at the end of each paragraph. In sentence mode you will get an <Enter> at the end of each sentence, giving a text which is in effect a long list of sentences.

Display: Normal/Accents etc. as Codes

in normal text format, or with accented characters shown as codes e.g. **caf<233> au lait** for **café au lait**.

Ignored tags: cut/visible

Your [Tags To Ignore](#) can be made visible or invisible.

Format: Plain Text/HTML/SGML

The default is [plain text](#) but if your text is in [HTML, XML or SGML](#) format you can change this setting and will see any [tags](#) separately in the left column.

Header: cut/visible

If your text is [tagged](#) and has a header, this too can be shown or cut out.

11.7 sentence 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 you can mark them with , then press . You will be told if you're in danger of jumbling up your text!

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

"It
is
good!"
Mary
wanted
more.

This will insert the following words (*Mary wanted more.*) into a new line below.

See also: [Viewer contents](#)

11.8 settings

The settings are standard ones found in most of the Tools:

[Colours](#)

[Font](#)

[Printing](#)

[Text Characteristics](#)

[Review all Settings](#)

11.9 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** is: a sentence ends if a full-stop, question-mark or exclamation-mark (.?!) is immediately followed by one or more [word separators](#) and if the next non-punctuation symbol is a capital letter A..Z or an accented capital letter, a number or a currency symbol. The same routine is used in **WordList**, though **WordList** attempts to distinguish between sentences and headings, so numbers of sentences in the two Tools are not likely to match.

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 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 letter (e.g. in *Wild Bill, the coachman, said.*)

Viewer 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](#) any odd one-word sentences.

How long is a sentence?

The storage space for each sentence or paragraph is 10,000 characters. **Viewer** can show up to 16,368 sentences or paragraphs. (If you estimate a conservative 10 words per sentence that's 160,000 words.)

Disk handling, Accents, etc.

For best results, use [ascii](#) or [ansi](#) versions of your two texts.

The joint text will be saved using a Windows [character set](#).

See also: Viewer contents

11.10 translation mis-matches

Viewer 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 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](#) or [join](#) and/or edit the text as appropriate, then save it.

See also: [The technical side...](#), [Finding unusual sentences](#), Viewer contents

11.11 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. **Viewer** will re-read the text file up to its limit of 16,368 sentences.

Don't like the colours

Change colours in the top-level *Settings* menu. The colours 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 contents

11.12 unusual sentences

Because **Viewer** 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!").

It can be useful to seek unusually short sentences to see whether your originals have been handled as you want.

This function also finds 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

F8 or *Find Short Lines*. **Viewer** will go to the next possibly problematic sentence or paragraph and you will probably want to [join](#) it by dragging it to the one above.

See also: [The technical side...](#), [Finding translation mis-matches](#), Viewer contents

WordSmith Tools

Reference

Section

XIII

12 Reference

12.1 32-bit version

This version of **Oxford WordSmith Tools** is a complete re-write in comparison to the earlier 16-bit versions, with lots of changes "under the hood". Some of the changes you will see are:

- long [filenames](#)
- better [tag and entity](#) handling including [Tag Concordancing](#)
- previous work can still be used, but it should be re-saved in the 32-bit format. You will get a suggestion to "Update" a data file if it is still in the old format.
- [zip file handling](#)
- easier exporting of data to Microsoft Word and [Excel](#)
- Unicode text handling, allowing more [languages](#) to be processed
- possibility of [altering the data](#) as it comes in, e.g. for language-specific lemmatisation
- the old limitations of 16,000 lines of data have gone. (The theoretical limit for a list of data is over 134 million lines.)

See also: [Contact Addresses](#).

12.2 acknowledgements

Oxford 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. **Concord** has a lot of additional features in this Windows version and all the code has been re-written, but the essential features of the design were there in *MicroConcord*.

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 6 and still using time-critical sections in Assembler. I am grateful to generations of students and colleagues at the [Department of English](#), University of Liverpool, and the MA Programme in Applied Linguistics at the Catholic University of São Paulo, for their feedback on aspects of the suite (including bugs!), and suggestions as to features it should have. 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](#) with any further ideas for developing **WordSmith Tools**.

12.3 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.
- 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.
- 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.

12.4 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 get one of these in any of the programs, you will 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 `wshell.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:



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 Oxford WordSmith Tools [Controller](#), retrieve your [saved results](#) from disk, and resume. If that doesn't work, try quitting Oxford WordSmith Tools overall, or quit Windows and then start it up again.

When you do quit, you should get this:



If you choose Yes, your email program will start up and a text will appear in it containing information which I can use to try to correct the problem in future. ***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](#)

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 [filename](#) or a number containing a comma.

See also: [logging](#), [troubleshooting](#).

12.5 Character Sets

12.5.1 overview

You need "[plain text](#)" in WordSmith. Not Microsoft Word .doc files -- which contain text and a whole lot of other things too that you cannot normally see.

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 wasteful of disk and memory space. So, at the time of writing, 1-byte character sets are still in very widespread use.

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 [define the character set in advance](#), 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](#) and View the file in question. While viewing you can change Text Characteristics until it looks right. 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 and [save it as text](#) again as a plain text file in Windows format, which is more up-to-date than DOS formats.

See also: [Choosing Accents & Symbols](#), [Accented characters](#); [Choosing Language](#)

12.5.2 accents & symbols

This feature of [Concordance Settings](#) and other dialogue boxes enables you to insert symbols and accented characters into your search-word, exclusion word or context word, etc. Just choose the symbol and drag it with your mouse to the place where you need it, or double-click on it.

You will see all the accented characters available, based on current settings such as the [character set](#) and [text characteristics](#), and a number of other symbols such as curly apostrophes. Some of these symbols (®,©,™ and curly apostrophes) may only be sought if your texts were saved in Windows [ANSI](#) format.

See also: [Choosing Language](#)

12.5.3 ansi and ascii

ASCII text, ANSI text, Text Only and DOS text are all names for plain text.

Most word-processors insert special hidden codes into text files to help them keep track of page numbers, bold type and so on. Oxford WordSmith Tools can handle them anyway but you'll get cleaner results if you use plain text without the hidden codes.

If your source texts were saved as "Text Only with line breaks" there will probably be one `<Enter>` every 70 or 80 characters at the end of each text line. If they were saved as "Text Only", the `<Enters>` will be equivalent to paragraph breaks. I recommend saving as "Text Only".

The Windows program **Notepad** (*Start | Program Files | Accessories*) makes plain text files or `.txt` files. It uses basic character sets e.g. **A** to **Z**, numbers and common punctuation symbols. The main difference is in the accented characters. For more on this, see [character sets](#).

See also : [HTML,SGML & XML](#).

12.5.4 DOS

DOS (text format before Windows) offered a range of character sets called "codepages". They all shared the same codes for the standard English alphabet (**a**, for example is always code 97) and common punctuation symbols, but included varying symbols for box-drawing, foreign language accents, etc.

If you process texts in German, Spanish, Russian, Greek, Polish, etc. you may need to find out which codepage was used when the texts were originally typed.

For example, the character **ã** is coded one way in codepage 850 (Multilingual) but differently in codepage 860 (Portuguese). It is simply not available at all in codepage 437 (the default codepage in the UK and USA). To alter or examine codepages, see a DOS manual or check the topic out on the web.

When it loads up, **Oxford WordSmith Tools** detects the current DOS code-page, so the codepage is only likely to need altering if you are using texts produced when another codepage was in use.

12.5.5 Windows

Windows character set codes are different from those in DOS or Unicode. (The £ symbol is code 156 in DOS but 163 in Windows.) In Windows 95 or later you can get non-Western fonts enabled via Microsoft Plus. If your texts were written using a Windows word-processor and [saved as text](#) in Windows, the accented characters will obey the Windows codes. You will have access to a few more symbols than in DOS (e.g. ®, ©, ™ and curly apostrophes).

Windows **Western** (1252) format includes:

Anglo-Saxon, Basque, Catalan, Danish, Dutch, English, Middle English, Finnish, French, German, Icelandic, Italian, Norwegian, Old Norse, Portuguese, Spanish, Swedish

Windows **Baltic** (1257) format includes:

Estonian, Latvian, Lithuanian

Windows **Central European** (1250) format includes:

Albanian, Bosnian, Croatian, Czech, Hungarian, Polish, Romanian, Serbian, Slovak, Slovene, Upper Sorbian, Lower Sorbian

Windows **Cyrillic** (1251) format includes:

Byelorussian, Bulgarian, Macedonian, Russian, Serbian (1251), Ukrainian

Windows **Greek** (1253) handles Greek

and Windows **Turkish** (1254) handles Turkish (what else?)

12.5.6 Unicode

A text format standard which uses 2 "bytes" per character. This allows for over 65,000 different characters and symbols to be displayed and makes it possible to show Chinese, Japanese, Cherokee and a whole lot of other languages.

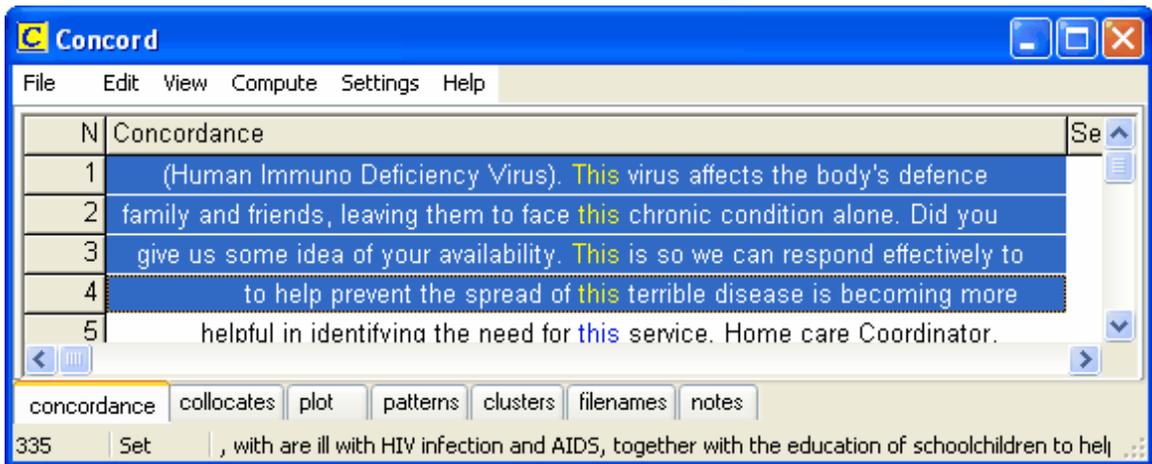
When choosing texts, you can press a button to test whether text files are encoded in Unicode.

12.6 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](#) (or [printing](#) to a file) and can also handle any graphic marks.

Example

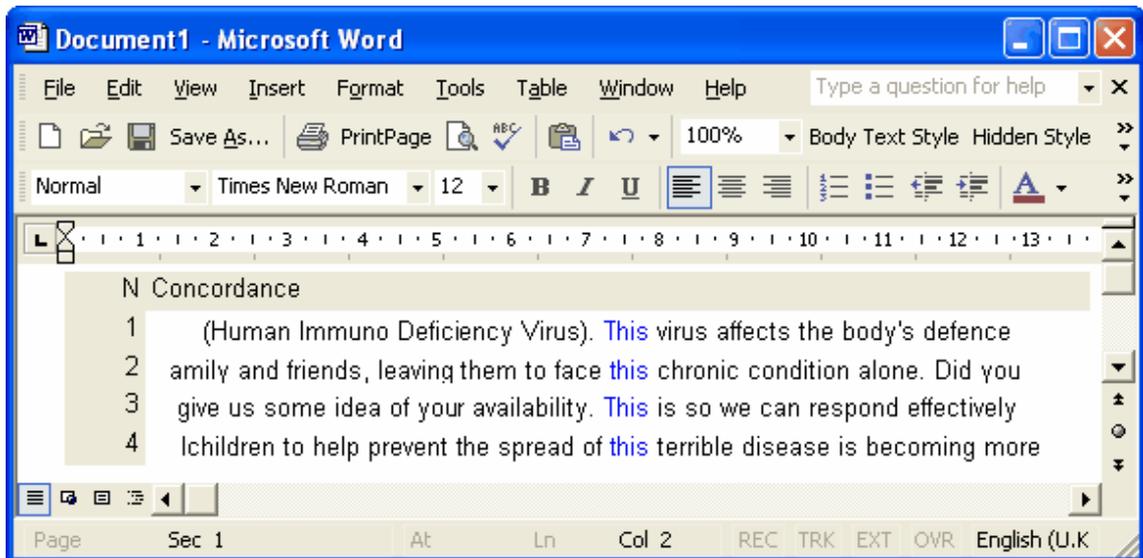
1. Select some data. Here I have selected the first 4 lines (of 335) of a concordance, just the visible text, no Set or Filenames information.



2. Hold down Control and press Ins or C. The data is now in the Windows "clipboard" ready for pasting into any other application, such as Excel, Word, Notepad, etc.

The data is automatically placed in the Clipboard in two different formats:

as a Picture



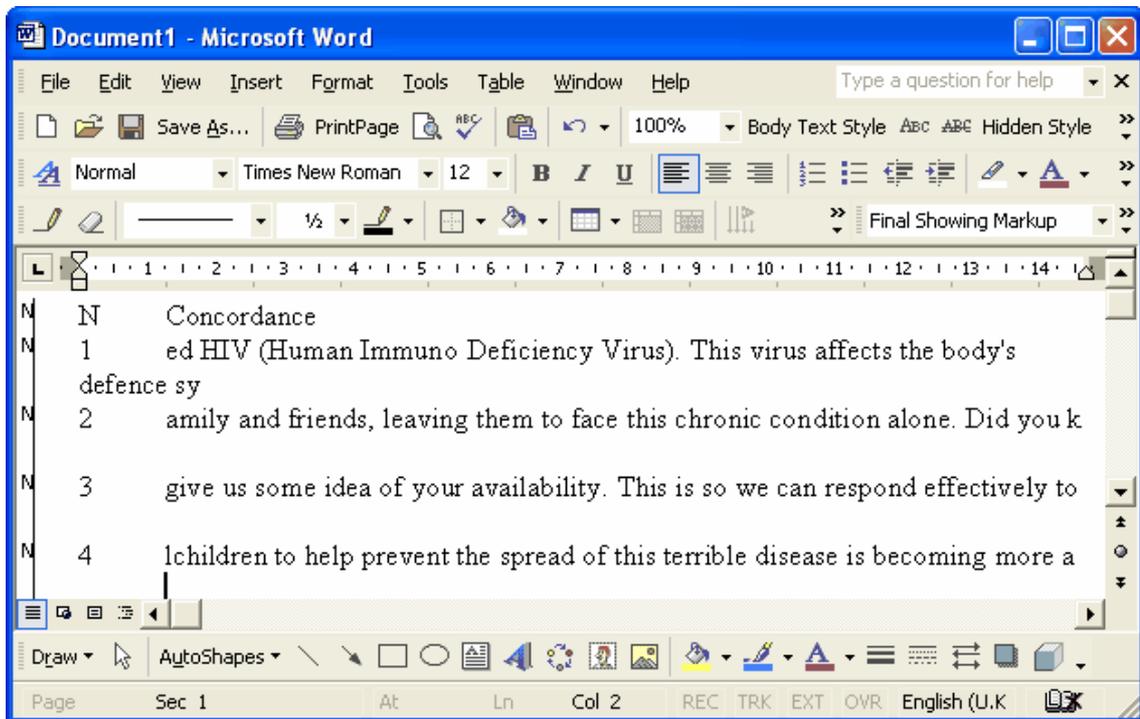
You will probably use this format for your dissertation and will have to in the case of plotted data. In this concordance, you get only the words visible in your concordance line (not the whole line). This is a graphic which includes screen colours and graphic data. If you subsequently click on the graphic 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!).

To get this, in Word, I chose *Edit | Paste Special | Picture (Enhanced Metafile)*. What you see in Word is very like what you see in Concord.

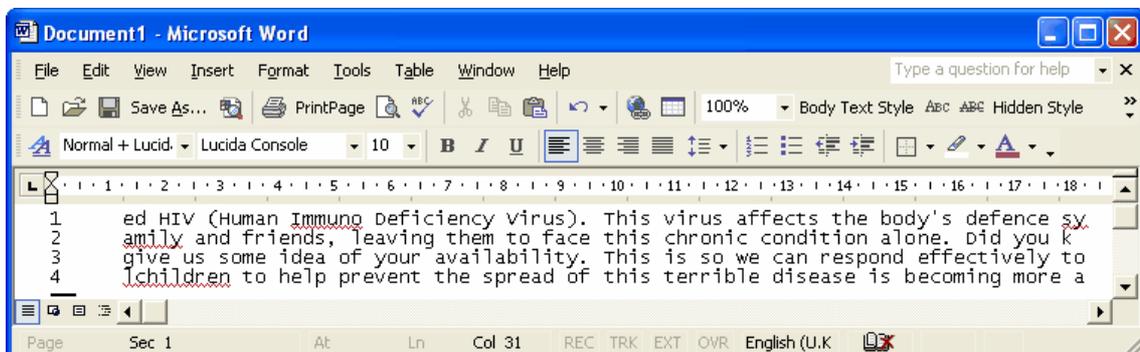
as plain text

Alternatively, you might want to paste as plain 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](#)™ (which will be better if you have graphic data, such as [Concord dispersion plots](#) or [KeyWords plots](#)).

Here the concordance or other data is 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 set in the [settings](#) for save as text, the default being 80.



Here, the concordance lines are copied, but of course they don't line up very nicely and it's hard to see the search-word (*this*). For the search-word to line up nicely, you should use a non proportional font, such as Courier or Lucinda Console, and it'll look like this.



Notice that at 10 point text in Lucida Console, the width of the text with 80 characters and the numbers at the left comes to over 18 cm. To avoid word-wrapping, I set the page format in Word to landscape. An alternative is to reduce the number of characters per line to say 50 or 60.

12.7 contact addresses

Downloads

You can get a more recent version at [my website](#). There are also some free extra downloads (programs, word lists, etc.) there too. And links to sources of free text corpora. The latest official [Oxford University Press](#) version is usually not as up to date.

Screenshots

visit [WordSmith 3 screenshots page](#) or <http://www.lexically.net/wordsmith/version4/screenshots/index.html> for screenshots of what Oxford 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://www.lexically.net/wordsmith/purchasing.htm> for details of suppliers.

Complaints & Suggestions

If you do not have the official OUP version but one from my website, please do not email OUP but me (Mike.Scott@liv.ac.uk). 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...

12.8 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?

12.9 Definitions

12.9.1 definitions

words

The word is defined as a *sequence of valid characters with a [word separator](#) at each end*. Valid characters include all the letters from A to Z, plus all accented characters which can be used in the current [character set](#), plus any user-defined acceptable characters to be included within a word (such as the apostrophe or [hyphen](#)).

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](#).

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](#) or [Concord cluster processing](#) there can be no certainty of this, though clusters often do match phrases or idioms. See also: [general cluster information](#).

sentences

The sentence is defined as *the full-stop, question-mark or exclamation-mark (.?!) immediately followed by one or more [word separators](#) and then a capital letter in the current language, a*

number or a currency symbol. (For more discussion see [Starts and Ends of Text Segments](#) or [Viewer technical information](#).)

paragraphs

Paragraphs are user-defined. See [Starts and Ends of Text Segments](#) for further details.

headings

Headings are also user-defined -- see [Starts and Ends of Text Segments](#).

See also: [Setting Text Characteristics](#), [Key-ness](#), [Key key-word](#), [Associate](#)

12.9.2 word separators

Conventionally one assumes that one word is distinguished from the next by the presence of spaces at either end. But **Oxford 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](#) 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.

12.10 demonstration version

The demonstration version of **Oxford WordSmith Tools** offers *all* the facilities of the complete suite, except that any screen which shows a list (of words in a wordlist, 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.)

Updating

To update your demo version, visit <http://www.lexically.net/wordsmith/purchasing.htm> for details of suppliers.

Upon receipt of the registration code, run **Oxford WordSmith Tools** as usual. You'll see a menu option called *Update from Demo*. Choose this and type in your registration name plus "any other details" such as your place of work, and the 20-letter registration code. The name and "any other details" must be at least 8 letters altogether.

Copy or paste the *exact* name which you supplied as your registered name for **Oxford WordSmith Tools** when you purchased the code. This name appears in the main window and whenever you access the *About* menu option (F9). Your software will then be fully enabled, and the *Update from Demo* menu option will disappear. (The `updater.exe` program will still be there in your `\wsmith4` folder, and can be used if you ever need to re-register.)

For Use on Network: check this if you are installing on a network drive and plan for users to access it from other PCs connected to the network. This can only be done if your licence permits it (not a single user licence).

Put an Icon on Desktop: check this if you want an icon for WordSmith on your desktop so you can access it easily.

Associate WordSmith files with Program: check this if you want your PC to know how to open files created with Oxford WordSmith Tools. (A `.cnc` file is to be opened using Concord, a `.lst` file with WordList, etc.) You may need Administrator rights to do this.

If you make a mistake and your registration fails, you can try again. If your registration succeeds but you decide to change the "any other details", run **updater.exe**, which you'll find in your **\wsmith4** folder.

See also: [Version Information](#), [Contact Addresses](#).

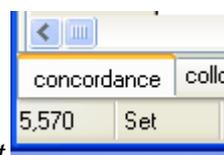
12.11 edit v. type-in mode

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](#) for the data, or for [blanking](#) out the search word.

In type-in mode, your key-presses are supposed to help you [get quickly](#) 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 by right-clicking on the word `Set` and choosing from

the  menu which opens up.

See also: [user-defined categories](#).

12.12 finding source texts

For some calculations the original source texts need to be available. For example, for Concord to show you [more context](#) than has been saved for each line, it'll need to re-read the source text. For KeyWords to calculate a [dispersion plot](#), 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](#).

If you have moved or deleted the source file(s) in the meantime, this won't be possible.

See also : [Editing filenames](#), [Choosing source files](#).

12.13 folders\directories

Found in main Settings menu in all Tools. Default folders can be altered in Oxford WordSmith Tools or set as [defaults](#) in `wshell.ini`.

- Concordance Folder: for your concordance files.
- KeyWords Folder: for your key-word list files.
- WordList Folder: where you will usually [save](#) your word-list files.
- Texts Folder: where your text files are to be found.

- Downloaded Media: where your [sound & video files](#) 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, Oxford 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](#) 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 names 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.

Tip

Use different folders for the different functions in Oxford 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](#) 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](#).

12.14 HistoryList

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.

12.15 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 (é corresponds to é)

See also: [Overview of Tags](#)

12.16 hyphens

The character used to separate words. The item "self-help" can be considered as 2 words or 1 word, depending on Text Settings.

12.17 Internet

Available under Adjust Settings | General in WordSmith Controller. Specify the full filename of your Internet browser or browse to find it. WordSmith will be able to use this to access help in the form of .htm files and screenshots included with the software.

12.18 limitations

The programs in **Oxford 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](#) 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](#) 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](#), 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](#) 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](#)

12.19 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](#), though you can specify an unlimited number of concordance search-words in a [search-word file](#).

Each concordance can store an unlimited number of collocates with a maximum [horizon](#) of 25 words to left and right of your search-word.

WordList limitations

A head entry can hold thousands of [lemmas](#), but you can only join up to 20 items in one go using F4. Repeat as needed.

[Detailed Consistency](#) 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](#)-windows per key-word [plot](#) display: 20.

number of windows of [associates](#) per key key-word display: 20.

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.

Text Viewer limitations

If you choose the View option  when choosing texts, **Viewer** 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** 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.
 You can only work on one dual-aligned text at a time.

See also: [General Limitations](#)

12.20 links between tools

Linkage with Word Processors, Spreadsheets, Browser etc.

All the windows showing lists or texts can easily copy selected information to the [clipboard](#). (Use Ctrl+Ins to insert).

 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://www.lexically.net>) you can click to access your browser.

Links between the various Tools

The programs in **Oxford WordSmith Tools** are linked to each other via [wshell.exe](#) (the one which says "Oxford WordSmith Tools [Controller](#)" in its caption, and is found in the top-left corner of your screen). This handles all the [defaults](#), such as colours, folders, fonts, stop lists, etc.

In general, if you press Control-C in **WordList** or **KeyWords** you'll go straight to a concordance, computed using the current word and using the current files.
 Press Control-W in **Concord** or **KeyWords** to start a wordlist 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](#), or view a dispersion plot.

Example: after computing a key words [database](#) 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.

12.21 list of 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.

-  **advice**
opens a window showing a map of Oxford WordSmith Tools, giving a view of where you are now and where you might go next; also offers advice depending on the Tool.
-  **associates**
opens a new window showing [Associates](#).
-  **auto-join**
joins ([lemmatises](#)) 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.
-  **clumps**
computes [clumps](#) in a keywords database
-  **regroup clumps**
[regroups](#) the clumps
-  **clusters**
computes concordance [clusters](#).
-  **collocates**
shows [collocates](#) using concordance data.
-  **compute**
calculates a [new column of data](#) 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](#) of numerical data.
-  **concord**
within KeyWords, WordList, starts Concord and concordances the highlighted word(s) using the original source text(s).
-  **copy**
allows you to [copy](#) your data to a variety of different places (the printer, [a text file](#), the [clipboard](#), etc.).
-  **double columns**
allows you to double the number of columns, so as to save paper when printing.
-  **edit**
allows [editing](#) of a list or searches for a word ([type-in search](#)).
-  **edit or type-in mode**
alternates between edit and type-in mode.
-  **filenames**
opens a new window showing the [filenames](#) from which the current data derived. If necessary you can [edit them](#).
-  **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 (also 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](#)).

 **layout**

This allows you to alter many settings for the [layout](#): the colour of each column, whether to hide a column of data, typefaces and column widths.

 **links**

computes [links](#) between words in a key-words plot.

 **mark**

marks an entry for [joining](#) or [finding files](#).

 **match lemmas**

checks each item in the list against ones from a text file of lemmatised forms and [joins](#) any that match.

 **match list**

matches up the entries in the current list against ones in a "[match list file](#)" or [template](#), marking any found with (~).

 **mutual information**

computes [mutual information](#) scores in a [WordList index list](#).

 **new...**

[gets you started](#) in the various Tools, e.g. to make a concordance, a word list, or a key words list.

open...

gives you a chance to choose a set of saved results.

 **patterns**

computes collocation [patterns](#).

 **play media**

plays a [media file](#).

 **plot**

opens a new window showing a [Concord dispersion plot](#) or [KeyWords plot](#).

 **print (also F3)**

previews your window data for printing; can print to file, which is equivalent to "[save as text](#)".

 **refresh**

re-reads your text file (in Viewer) or re-draws the screen (in Print Preview).

 **replace**

search & replace, e.g. to replace drive or folder data, when [editing file-names](#) where the source texts have been moved.

 **re-sort**

re-sorts lists (e.g. in frequency as opposed to alphabetical order) in [Concord](#), [KeyWords](#) or [WordList](#).

 **ruler**

shows/hides vertical divisions in any list; text divisions in a [KeyWords plot](#). Click ruler in a menu to turn on or off or change the number of ruler divisions for a [plot](#).

 **save (also F2)**

[saves your data](#) 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. This is not the same as saving your current data so as to get it back later within Oxford WordSmith Tools.

-  **search (also F12)**
[searches](#) within a list.
-  **shrink**
 reduces the height of all rows to a smaller fixed height. See grow () above.
-  **skim**
 in Viewer, allows timed skimming through a text.
-  **statistics**
 opens a new window showing [detailed statistics](#).
- 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](#), e.g. proportion of lemmas to word-types.
-  **swap columns for rows**
 swaps the columns and rows. WordList [statistics](#) are shown by default with the file data in each column. Click this button to swap the row data with the column data.
- toolbar**
 toggles on & off a toolbar with the same buttons on it as the ones you chose when you [customised popup menus](#).
-  **unjoin**
 unjoins any entries that have been joined, e.g. [lemmatised](#) entries.
-  **view source text**
[shows the source text](#) and highlights any words currently selected in the list.
-  **Microsoft Word™**
 sends formatted data to Word.
-  **wordlist**
 within KeyWords, [makes a word list](#) using the current data.
-  **zap**
[zaps](#) any deleted entries.

see also: [Keyboard Shortcuts](#), [Customising popup menus](#).

12.22 keyboard shortcuts

scrolling windows:

Control-Home to top of scrollable list

Control-End to last line of list

if it's ordered alphabetically, [type-in your search-word](#)

and if it scrolls horizontally:

Home to left edge

End to right edge

Control-Right one word to right

Control-Left one word to left

hotkeys:

Ctrl-C [call](#) **Concord** from within another Tool

Ctrl-W [call](#) **WordList** from within another Tool

Ctrl-Ins copy blocked section to [clipboard](#)

Shift-cursor keys block a section

F1	help 
F2	save results 
F3	print results 
F4	join entries 
F5	mark entries for joining 
F6	re- sort 
Ctrl/F6	reverse word sort 
F7	view source text 
F8	seek short sentences (Viewer)
F9	About box (shows version-date and memory availability)
F12	search within a list 
Ctrl/M	Merge 2 word lists or KeyWords databases
Alt/H	access to H elp sub-menus
Alt/W	access to S ettings sub-menus
Alt/X	access to W indow sub-menus
Alt/X	e X it the Tool
Ctrl/Z	Zap deleted lines

see also: [Menu items and Buttons](#)

12.23 long file names

This version of **WordSmith** handles long filenames correctly.

12.24 machine requirements

This version of **Oxford WordSmith Tools** is designed for machines with:

- at least 256MB of RAM (you might be OK with 128 but probably not on Windows XP)
- at least 40MB of hard disk space
- Windows™ 98, NT, 2000, XP or later, or an emulator of one of these if using an Apple Mac or Unix system. It may also work on Windows 95 2nd edition, I don't know...

You will find it runs better on a [faster](#) machine, especially if there's plenty of [RAM](#).

12.25 manual for WordSmith Tools

This help file exists in the form of a manual, which you get when you [install](#). The file (**wshell.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: check out [Contact Addresses](#).

12.26 numbers

Depending on Text Settings, you might wish to include or exclude numbers from word lists.

12.27 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](#) (1998: 190-191), which he reports as having been evaluated as the most reliable.

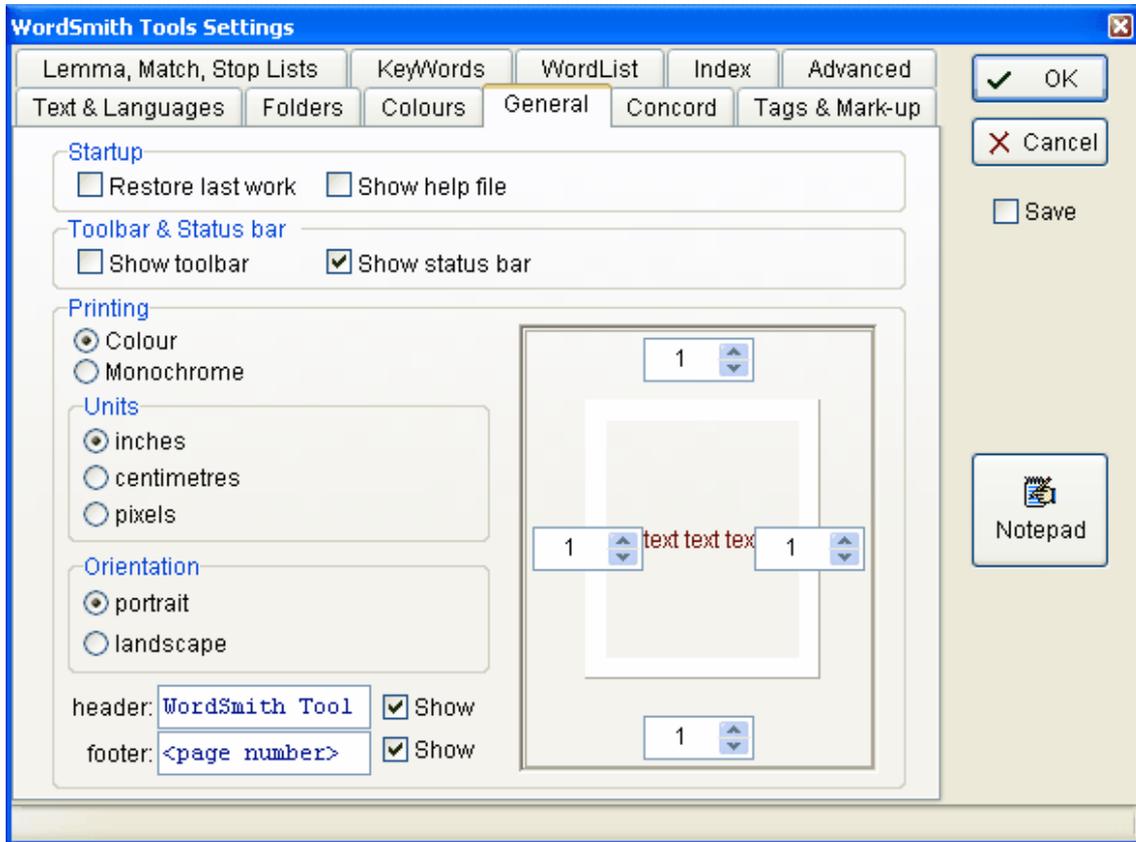
Like the [ruler](#), it divides the plot into 8 segments for this.

It ranges from 0 to 1, with 0.9 or 1 suggesting very uniform dispersion and 0 or 0.1 suggesting "burstiness" ([Katz](#), 1996)

See also: [KeyWords plot](#), [Concord dispersion plot](#).

12.28 printer settings

Found in *Adjust Settings | General* in the Oxford WordSmith Tools [Controller](#).



Startup

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

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.

Printing

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](#). 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. You can also setup your printer in [Print Preview](#).

Confirmation

You can set Oxford WordSmith Tools to confirm a print job in the [defaults \(wshell.ini\)](#) file. If this contains the line `confirm printing=YES` then every time you print you'll be told which lines of the current concordance or list were printed.,

See also : [Printing](#)

12.29 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 **Oxford 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](#), 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 64MB of RAM or more you should be able to have a copy of a wordlist 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](#)

12.30 reference corpus

Reference Corpus

A corpus of text which you use for comparative purposes. For example, you might want to compare a piece of text with the [British National Corpus](#), a collection of 100 million words. Useful when computing [key words](#).

In the [Controller](#) you can [set your reference corpus wordlist](#) for KeyWords and Concord to make use of.

12.31 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 **Oxford 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 **Wshell** menu option or permanently in `wshell.ini` (in your \wsmith4 folder).

12.32 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 calls them "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](#), and semantic prosody (the tendency for **cause** to come with negative effects such as **accident**, **trouble**, etc.).

Oxford WordSmith Tools gives you two opportunities for identifying word clusters, in [WordList](#) and [Concord](#). They use different methods. Concord only processes concordance lines, while WordList processes whole texts.

How Concord does it...

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.)

12.33 speed

To make a wordlist on 4.2 million words used to take about 20 minutes on a 1993 vintage 486-33 with 8Mb of [RAM](#). 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](#) 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 **Oxford 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](#) a document at the same time should be fine.

12.34 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](#).

12.35 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 **Oxford 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. **Oxford WordSmith Tools**, as their 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](#).

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.

Oxford 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 **Oxford 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 : [Acknowledgements](#)

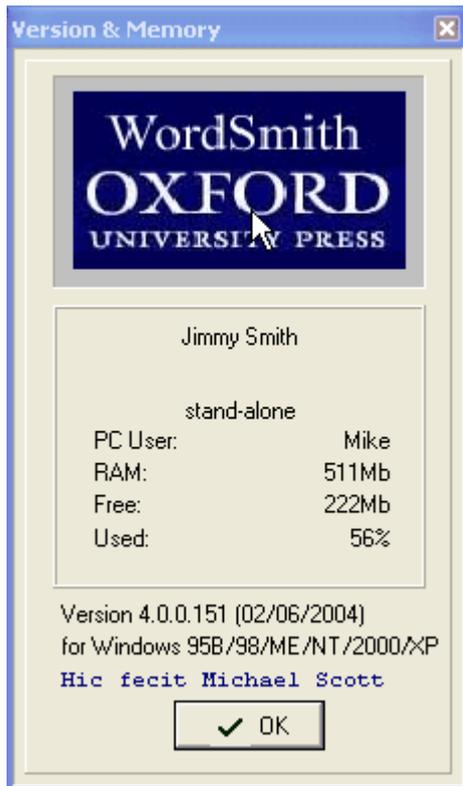
12.36 version information

This help file is for the current version of **Oxford WordSmith Tools**.

The version of **Oxford WordSmith Tools** is displayed in the *About* option (F9) which also shows your registered name and the amount of [memory](#) 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 Oxford graphic in the About box to see your current code.



See also: [32-bit Version Differences](#), [Demonstration Version](#), [Contact Addresses](#).

12.37 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*TM and *Winzip*TM. 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.

WordSmith Tools

Troubleshooting

Section

XIII

13 Troubleshooting

13.1 list of FAQs

See also: [logging](#).

These are the Frequently Asked Questions.

There's a much longer list of explanations under [Error Messages](#).

[Can't process apostrophes](#)

[Is this Russian, Greek or English? strange symbols in display](#)

[It crashed](#)

[It doesn't even start!](#)

[It takes ages!](#)

[Keys don't respond](#)

[Line beyond demo limit](#)

[Mismatch between Concord and WordList results](#)

[No tags visible in concordance](#)

[Printing problem](#)

[Text unreadable because of the colours](#)

[Too much or too little space between columns](#)

[Wordlist out of order](#)

[Won't slice pineapples](#)

13.2 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: drag the symbol from the set below when entering your [search word](#), or else replace them in your text files using Text Converter.

See also: [settings](#)

13.3 column spacing

column spacing is wrong

You can alter this by clicking on the [layout](#) button.

13.4 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](#) 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](#).

Solution: specify the tag file and run the concordance again.

13.5 Concord/WordList mismatch

Concord/WordList mismatch

If **WordList** finds a certain number of occurrences of a [\(word list\) cluster](#) 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](#) it will take punctuation into account.

13.6 crashed

it crashed!

Solution: quit **Oxford WordSmith Tools** and enter again. If that fails, quit Windows and try again.

13.7 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 Oxford WordSmith Tools window, and choose *Settings | Register* in the menu.

If you haven't got the 20-character registration code, contact [Oxford University Press](#). The **only** difference between a [demonstration version](#) 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.

13.8 funny symbols

weird symbols

funny symbols when using Oxford WordSmith Tools

1. Check your text files. Read them in **Notepad**. Do they contain lots of strange symbols? These may be hidden codes used by your usual word-processor. Solution: read them into your usual word-processor and *Save As*, with a new name, in [plain text](#) format, sometimes called "Text Only" or **.txt**.
2. *Choose Texts*, highlight the text file, and before pressing *OK*, press *View*. Does it contain strange symbols? Solution: change *Text Settings*; try going from one of the DOS character sets to Windows or vice-versa. The text was clean [ASCII](#) but **Oxford WordSmith Tools** thought it was Windows ANSI.
3. Funny symbols in a word list may well also be caused by mis-spellings in the original text files.

Greek, Russian, etc.

4. If the text is in Russian, Greek, etc. you will need an appropriate font, obtainable e.g. for Windows 95 via Microsoft Plus.
5. If you have several lists open which use *different* character sets, and you change [Font](#) or [Text Characteristics](#), 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

Oxford WordSmith Tools can [Save](#) or *Save As* and [Saves as text](#)" by [printing](#) to a file. "Save" and "Save As" will store the file in a format for re-use by **Oxford WordSmith Tools**. 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](#).

13.9 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.

13.10 keys don't respond

Keys don't respond

If a key press does nothing, it is probably because the wrong 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 anywhere within the window which you want to use, then press the key you wanted.

13.11 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.)

13.12 printer didn't print

printing problem

If your printing comes out with one or more column blank but others printed correctly, you may have a printer which can only manage black and white and not shades of grey. In the [Controller](#), change the setting (*Adjust Settings | General*) to monochrome.

13.13 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](#) 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 **Oxford 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, not the network.

13.14 won't start

it doesn't even start

Yikes!

13.15 word list out of order

wordlist 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](#) and re-compute the wordlist.

WordSmith Tools

Error Messages

Section

XIV

14 Error Messages

14.1 list of error messages

List of Error Messages

See also: [Troubleshooting](#).

[Can only save WORDS as ASCII](#)

[Can't call other Tool](#)

[Can't make folder as that's an existing filename](#)

[Can't merge list](#)

[Can't read file](#)

[Character set reset to <x> to suit <language>](#)

[Concordance file is faulty](#)

[Concordance stop list file not found](#)

[Conversion file not found](#)

[Destination folder not found](#)

[Disk problem: File not saved](#)

[Dispersions go with concordances](#)

[Drive not valid](#)

[Failed to access Internet](#)

[Failed to create new folder name](#)

[File access denied](#)

[File contains none of the tags specified](#)

[File not found](#)

[Filenames must differ!](#)

[Full drive:\folder name needed](#)

[function not working properly yet](#)

[INI file not found](#)

[Invalid Concordance file](#)

[Invalid file name](#)

[Invalid Keywords Database file](#)

[Invalid Keywords file](#)

[Invalid Wordlist Comparison file](#)

[Invalid Wordlist file](#)

[Joining limit reached: join & try again](#)

[Key words file is faulty](#)

[Keywords Database file is faulty](#)

[Limit of 500 file-based search-words reached](#)

[Links between Tools disrupted](#)

[Match list details not specified](#)

[Must be a number](#)

[Network registration running elsewhere or vice-versa](#)

[No access to text file: in use elsewhere?](#)

[No associates found](#)

[No clumps identified](#)

[No clusters found](#)

[No collocates found](#)

[No concordance entries found](#)

[No concordance stop list words](#)

[No deleted lines to Zap](#)

[No entries in Keywords Database](#)

[No Key Words found](#)
[No key words to plot](#)
[No keyword stop list words](#)
[No lemma list words](#)
[No match list words](#)
[No room for computed variable](#)
[No statistics available](#)
[No stop list words](#)
[No such file\(s\) found](#)
[No tag list words](#)
[Not a valid number](#)
[No wordlists selected](#)
[Original text file needed but not found](#)
[Registration string is not correct](#)
[Registration string must be 20 letters long](#)
[Short of Memory!](#)
[Source Folder file\(s\) not found](#)
[Stop list file not found](#)
[Stop list file not read](#)
[Tag file not found](#)
[Tag list file not read](#)
[This function is not yet ready!](#)
[This is a demo version](#)
[This program needs Windows 95 or greater](#)
[To stop getting this annoying message, Update from Demo in setup.exe](#)
[Too many ignores \(50 limit\)](#)
[Too many sentences \(8000 limit\)](#)
[Two files needed](#)
[Truncating at xx words -- tag list file has more!](#)
[Unable to merge Keywords Databases](#)
[Why did my search fail?](#)
[Word list file not found](#)
[Wordlist comparison file is faulty](#)
[Word-list file is faulty](#)
[Oxford WordSmith Tools has expired: get another](#)
[Oxford WordSmith Tools already running](#)
[WordSmith version mis-match](#)
[xx days left](#)

14.2 .ini file not found

.ini file not found

On starting up, **WordSmith** looks for the **wshell.ini** file which holds your current [defaults](#). If you've removed or renamed it, restore it. This file should be in the same folder as the Tools are in.

14.3 base list error

base list error

WordSmith is trying to access a word or concordance line above or below the top or bottom of the data computed. This is a bug.

14.4 can only save words as ASCII

Can only save WORDS as Plain Text

Oxford 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](#) and pasting it into your word-processor.

14.5 can't call other tool

Can't call other Tool

Inter-Tool communication has got disrupted. [Save](#) your work, first. Then, if necessary, close down **Oxford WordSmith Tools** altogether, then start the main **wshell.exe** program again.

14.6 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.

14.7 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.

14.8 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.

14.9 can't read file

Can't read file

If this happens when starting up **Oxford WordSmith Tools**, there is probably a component file missing. One example is **sayings.txt**, which holds sayings that appear in the main [Controller](#) 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, 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.

14.10 character set reset to <x> to suit <language>

Character set reset to <x> to suit <language>

Prior to version 2.00.07, **Oxford WordSmith Tools** handled fewer [character sets](#) and [languages](#) 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](#), *Adjust Settings | Text*, then re-save the list.

14.11 concordance file is faulty

Concordance file is faulty

Each type of file created by **Oxford 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**.

14.12 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 [filename](#), remember to include the full drive and folder as well as the filename itself.

14.13 confirmation messages: okay to re-read

Okay to re-read?

A confirmation message. To proceed, **Viewer** 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** will try to read the whole text file (up to the limit of 16,368 sentences). If you have a very big file on a slow CD-ROM drive, this will take some time.

14.14 conversion file not found

Conversion file not found

You typed in the name of a non-existent file. If typing in a [filename](#), remember to include the full drive and folder as well as the filename itself.

14.15 destination folder not found

Destination folder not found

WordSmith couldn't find that folder; perhaps it's mis-spelt.

14.16 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.

14.17 dispersions go with concordances

Dispersions go with concordances

They can't be [saved](#) separately.

14.18 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:.

14.19 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.

14.20 failed to create new folder name

Failed to create new folder

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.

14.21 failed to read file

Failed to Read

This may have happened because your disk filing system has got screwed up. This is especially likely to occur if you often use large files in your word processor. I would recommend you to run *System Tools | Scandisk*.

14.22 failed to save file

Failed to Save

This may have happened 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. I would recommend you to run *System Tools | Scandisk*. If you're working on a network, you will be able to [save](#) on certain drives and folders but not others; the solution is to try again on drive A: or a hard disk drive which you have the right to save to.

14.23 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.

14.24 file contains none of the tags specified

File contains none of the tags specified

You specified tags, but none of them were found.

14.25 file has "holes"

File has "holes"

Your text file is defective. It may well contain useful text, but it also contains at least one unrecognised character 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](#). You will see the context where the problem occurred and will be told roughly how far into the text it was detected.

WordSmith can proceed if you wish but you get a chance to skip the text.

You can solve this problem -- which will come each time you choose that text file -- by reading the text file into a word processor and re-saving it as a plain .txt file. Also, in [File Utilities](#) there is a tool for finding such files.

14.26 file not found

File not found

This message, like [Original Text not found](#), 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 filenames](#) in the filename window () of this list.

Or the message may come after you've supplied the filename yourself. You may have mis-typed it. Is it a Windows 95 or NT [long filename](#)?. If typing in a [filename](#), remember to include the full drive and folder as well as the filename itself.

14.27 filenames must differ!

Filenames must differ

You can't compare a file with itself.

14.28 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.

14.29 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.

14.30 full drive & folder name needed

Full drive:\folder name needed

When typing in a [filename](#), remember to include the full drive and folder as well as the filename itself.

14.31 function not working properly yet

function not working properly yet

This is a function under development, still not fully implemented.

14.32 invalid concordance file

Invalid Concordance file

Each type of file created by **Oxford 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**.

14.33 invalid file name

Invalid file name

[Filenames](#) may not contain spaces or certain symbols such as **?** and *****. In Windows before Windows 95 they had to be restricted to 8 letters and a dot and three more, too. Try again.

14.34 invalid KeyWords database file

Invalid Keywords Database file

Each type of file created by **Oxford 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**.

14.35 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 and both must be sorted in the same way (alphabetical order, ascending). 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. Check they have both been made with the same language settings and if necessary re-compute one or both of them.

14.36 invalid WordList comparison file

Invalid Wordlist Comparison file

Each type of file created by **Oxford 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**.

14.37 invalid WordList file

Invalid Wordlist file

Each type of file created by **Oxford 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**.

14.38 joining limit reached

Joining limit reached: join & try again

Only a certain number of words can be [lemmatised](#) 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.

14.39 KeyWords database file is faulty

Keywords Database file is faulty

Each type of file created by **Oxford 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**.

14.40 KeyWords file is faulty

Key words file is faulty

Each type of file created by **Oxford 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**.

14.41 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](#) to tell **Concord** to do them in a batch, where the limit is 500.

14.42 links between Tools disrupted

Links between Tools disrupted

Oxford WordSmith Tools Controller 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](#) are vital in this suite, you should exit **WordSmith** and re-enter.

14.43 match list details not specified

Match list details not specified

You pressed the [Match List](#) button but then failed to choose a valid match list file or else to type in a template for filtering. Try again.

14.44 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).

14.45 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 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.

14.46 network registration used elsewhere

Network registration running elsewhere or vice-versa

The 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.

14.47 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.

14.48 no associates found

No associates found

Alter settings (*Settings | Min & Max Frequencies*) and try again.

14.49 no clumps identified

No clumps identified

Alter settings and try again.

14.50 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.

14.51 no collocates found

No collocates found

In the [Controller](#), alter the settings (*Adjust Settings | Concord | Min. Frequency*) and try again. There were too few concordance lines to find the minimum number needed.

14.52 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](#), check the format of your texts. If you're using a [search-word file](#), ensure this was prepared using a plain Windows word-processor such as Notepad.

Have you specified any [wildcards](#) (* 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](#) or two. You're more likely to find book* than book.

14.53 no concordance stop list words

No concordance stop list words

14.54 no deleted lines to zap

No deleted lines to Zap

You pressed Alt-Z but hadn't any deleted lines to [zap](#). No harm done.

14.55 no entries in KeyWords database

No entries in Keywords Database

Alter settings and try again.

14.56 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](#) too small for any key words to be detected. For very short texts a minimum frequency of 2 may be needed.

14.57 no key words to plot

No key words to plot

Had you deleted them all?

14.58 no Keywords stop list words

No keyword stop list words

WordSmith either failed to read your stop-list file or it was empty.

14.59 no lemma list words

No lemma match list words

WordSmith either failed to read your lemma list file or it was empty.

14.60 no match list words

No match list words

WordSmith either failed to read your [match list](#) 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!

14.61 no room for computed variable

No room for computed variable

There isn't enough space for the variable you're trying to compute.

14.62 no statistics available

No statistics available

Some types of word list created by **Oxford 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.

14.63 no stop list words

No stop list words

WordSmith either failed to read your stop-list file or it was empty.

14.64 no such file(s) found

No such file(s) found

You typed in the name of a non-existent file. If typing in a [filename](#), remember to include the full drive and folder as well as the filename itself.

14.65 no tag list words

No tag list words

WordSmith either failed to read your tag file or it was empty.

14.66 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.

14.67 not a valid number

Not a valid number

Either you've just typed in, or else **Oxford WordSmith Tools** has just attempted to read (e.g. from **wshell.ini**, the [defaults](#) 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 **wshell.ini** file for mistakes.

14.68 not a WordSmith file

The file you are trying to open is not a Oxford WordSmith Tools file. WordSmith makes files containing your results, files whose names end in **.LST**, **.CNC**, **.XWS**, 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** 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](#)

14.69 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 4 and the format has had to change (sorry!)

If the former, you may be able to convert it using the [Converter](#).

14.70 nothing activated

Nothing activated

Some forms have choices labelled "Activated" which you can switch on and off. If they are un-checked, you can still see what they would be but **WordSmith** will ignore them.

14.71 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](#) 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!

14.72 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](#) 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.

14.73 registration code in wrong format

Registration code must be as in this example

2 letters or numbers, a dot, then 4 numbers, dot, 4 numbers etc.
Example: XX.1234.5678.9012.3456 (dots every 4 letters)

14.74 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](#) mode.

14.75 short of memory

Short of Memory!

An operation could not be completed because of shortage of [RAM](#)

14.76 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 [filename](#), remember to include the full drive and folder as well as the filename itself.

14.77 stop list file not found

Stop list file not found

You typed in the name of a non-existent file. If typing in a [filename](#), remember to include the full drive and folder as well as the filename itself.

14.78 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.

14.79 tag file not found

Tag File not found

You typed in the name of a non-existent file. If typing in a [filename](#), remember to include the full drive and folder as well as the filename itself.

14.80 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, 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.

14.81 this function is not yet ready

This function is not yet ready!

Temporary message, for functions which are still being tested.

14.82 this is a demo version

This is a demo version

You will probably want to [upgrade](#) to the full version.

14.83 this program needs Windows 98 or greater

This program needs Windows 98 or better

As of Version 4.0, this is a 32-bit program (and a 32-bit help file).

14.84 to stop getting this message ...

Get an update. This is "annoyware" for the [demonstration version](#).

14.85 too many requests to ignore matching clumps

The limit is 50. Do any remaining joining manually.

14.86 too many sentences

The limit is 8,000. Do the task in pieces.

14.87 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!

14.88 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.

14.89 unable to merge Keywords Databases

Perhaps there wasn't enough [RAM](#) to carry out the merge.

14.90 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.)

14.91 word list file is faulty

Each type of file created by **Oxford 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**.

14.92 word list file not found

You typed in the name of a non-existent file. If typing in a [filename](#), remember to include the full drive and folder as well as the filename itself.

14.93 WordList comparison file is faulty

Each type of file created by **Oxford 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**.

14.94 WordSmith Tools already running

Don't try to start **Oxford 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.)

14.95 WordSmith Tools expired

Message for limited period users only. Your version of **Oxford WordSmith Tools** has passed its validity and is now in [demo](#) mode. Download another from the [Internet](#).

14.96 WordSmith version mis-match

Since the various Tools are [linked](#) 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](#).

14.97 XX days left

Message for limited period users only. At the end of this time **WordSmith** will revert to [demo](#) mode.

Index

- . -

.ini files 48

- 2 -

25 lines 189

- 3 -

32-bit version 163

- A -

about option 174
 accents 166
 accents & symbols 166
 accents window 19
 accessing previous results 45
 accurate sort in WordList 129
 acknowledgements 163
 add to text 95
 adding notes to data 19
 adjust settings 19
 adjusting with mouse 157
 advanced concordance settings 79
 advanced settings 36
 aligning 157
 alignment 40
 altering your data 33
 alternative search words 89
 alt-tab 54
 ansi 166
 apostrophes in sorting 87
 ascii 166
 associate defined 94
 associated entries 113
 associates 95
 asterisk 89
 auto-joining lemmas 107
 autoload tag file 58

automated file-based concordancing 83

- B -

Babbage 190
 Baltic 23
 batch choosing 95
 batch processing 19
 batch processing: file-names 22
 batch processing: folders 22
 bibliography 163
 blanking out entries 67
 BNC handling of sentences and headings 64
 BNC Sampler version 171
 BNC: selecting between texts 61
 BNC: selecting within texts 62
 BNC: tag file 62
 BNC: text format 173
 boolean and/not 82
 boolean or 89
 bracket first line 145
 browsing original 77
 bugs 164
 burstiness 180
 buttons 176

- C -

calculating a plot 101
 calculation of KeyWords 99
 call a concordance 97
 calling other tools 175
 cannot compare word-lists in different languages 199
 can't see Concord tags 188
 case sensitivity 89
 CD-ROM version: defaults 48
 CD-ROM: speed 183
 CD-ROM: storage 182
 Central European 23
 changing colours 27
 changing font 22
 changing from edit to type-in mode 172
 character sets 165
 characters in save as text 90
 characters within a word 52

- Charles Babbage 190
 - check current version 17
 - Cherokee 167
 - Chinese 167
 - chi-square 99
 - Choose Languages: overview 6
 - choosing files from standard dialogue box 26
 - choosing texts 24
 - choosing your files 95
 - class instructions 40
 - clear previous selection 24
 - clipboard 167
 - clumps 95
 - clumps: regrouping 103
 - cluster: definition 170
 - clusters 183
 - clusters in KeyWords 96
 - cocoa tags 58
 - codepages 165
 - codes 165
 - collocates 72
 - collocates: display 70
 - collocates: horizons 69
 - collocates: minimum frequency 69
 - collocates: sorting 88
 - collocation associates 95
 - collocation patterns 86
 - collocation: settings 69
 - collocation: specifications 69
 - colours 27
 - colours in tags 62
 - column totals 29
 - column width 40
 - columns in printing 181
 - comparing wordlists 109
 - comparison display 110
 - compute new column of data 30
 - Concord: categories 68
 - Concord: clusters 73
 - Concord: collocation 72
 - Concord: creating exercises 67
 - Concord: limitations 174
 - Concord: multiple search-words 83
 - Concord: nearest tag 84
 - Concord: overview 5, 67
 - Concord: patterns 86
 - Concord: saving and printing 76
 - Concord: sorting 87
 - Concord: sound and video 76
 - Concord: source text file 77
 - Concord: starting tips 12
 - Concord: stretching the display to see more 77
 - Concord: text segments 88
 - Concord: uniform plot 74
 - Concord: viewing options 76
 - Concord: what you see and can do 77
 - Concord: wildcards 89
 - Concord: zapping unwanted lines 82
 - concordance batch processing 79
 - concordance display 77
 - concordance settings 79
 - concordancing on tags 81
 - confirmation messages: okay to re-read 196
 - consistency analysis (detailed) 111
 - consistency analysis (simple) 112
 - consistency lists: sorting 124
 - contact addresses 170
 - context word 82
 - controller (wshell.exe) 5
 - convert data from old version 135
 - converter 150
 - copy choices 32
 - copy data to Word 167
 - copy: all 32
 - copy: selective 32
 - copy: specify 32
 - couldn't merge KW databases 207
 - count data frequencies 30
 - creating a database 97
 - custom .dll file 33
 - custom processing 33
 - custom settings 48
 - custom settings for BNC tags 60
 - customising popup menu 36
 - cutting line starts 62
 - Cyrillic 23
- D -**
- data as text file 101
 - database construction 97
 - database statistics 99

date format 170
 decimal places 40
 decimal view 158
 defaults 48
 defining multimedia tags 59
 definition of associate 94
 definition of key key-word 98
 definition of key-ness 98
 definitions 170
 deleting entries 54
 demonstration version 171
 desktop 174
 Dickens text 24, 45
 directories 172
 disambiguation 95
 dispersion 74
 dispersion plot: sorting 88
 displaying comparisons 110
 DOS codes 166
 download new version 17
 dual-text aligning with Viewer 157

- E -

edit mode 172
 editing concordances 82
 editing WordList entries 38
 end of heading marker 52
 end of paragraph marker 52
 end of sentence marker 52
 end of text separator 145
 end-of-text symbols 146
 English 190
 entity references 57
 error messages 193
 error messages: .ini file not found 194
 error messages: base list error 194
 error messages: can only save words as ASCII 195
 error messages: can't call other tool 195
 error messages: can't make folder as that's an existing filename 195
 error messages: can't merge list with itself! 195
 error messages: can't read file 195
 error messages: character set reset to <x> to suit <language> 196
 error messages: concordance file is faulty 196
 error messages: concordance stop list file not found 196
 error messages: conversion file not found 196
 error messages: destination folder not found 196
 error messages: disk problem -- file not saved 196
 error messages: dispersions go with concordances 197
 error messages: drive not valid 197
 error messages: failed to access Internet 197
 error messages: failed to create new folder 197
 error messages: failed to read file 197
 error messages: failed to save 197
 error messages: file access denied 197
 error messages: file contains "holes" 198
 error messages: file contains none of the tags specified 198
 error messages: file not found 198
 error messages: filenames must differ 198
 error messages: form incomplete 198
 error messages: full drive & folder name needed 199
 error messages: function not working properly yet 199
 error messages: invalid concordance file 199
 error messages: invalid file name 199
 error messages: invalid KeyWords database file 199
 error messages: invalid KeyWords file 199
 error messages: invalid WordList comparison file 199
 error messages: invalid WordList file 200
 error messages: joining limit reached 200
 error messages: KeyWords database file is faulty 200
 error messages: KeyWords file is faulty 200
 error messages: limit of file-based search-words reached 200
 error messages: links between Tools disrupted 200
 error messages: match list 201
 error messages: must be a number 201
 error messages: network registration used elsewhere 201
 error messages: no access to text file - in use elsewhere? 201
 error messages: no associates found 201
 error messages: no clumps identified 201
 error messages: no clusters found 201

error messages: no collocates found 202
 error messages: no concordance entries found 202
 error messages: no concordance stop list words 202
 error messages: no deleted lines to zap 202
 error messages: no entries in KeyWords database 202
 error messages: no key words found 202
 error messages: no key words to plot 202
 error messages: no KeyWords stop list words 203
 error messages: no lemma list words 203
 error messages: no match list words 203
 error messages: no room for computed variable 203
 error messages: no statistics available 203
 error messages: no stop list words 203
 error messages: no such file(s) found 203
 error messages: no tag list words 203
 error messages: no word lists selected 204
 error messages: not a valid number 204
 error messages: not a WordSmith file 204
 error messages: nothing activated 204
 error messages: original text file needed but not found 204
 error messages: printer needed but not found 205
 error messages: registration string is not correct 205
 error messages: registration string must be 20 letters long 205
 error messages: short of memory 205
 error messages: source folder file(s) not found 205
 error messages: stop list file not found 205
 error messages: stop list file not read 205
 error messages: tag file not found 206
 error messages: tag file not read 206
 error messages: the program needs Windows 98 or greater 206
 error messages: this function is not yet ready 206
 error messages: this is a demo version 206
 example 97
 Excel 46
 exercises 67
 exiting 45
 expiry date 207
 export to spreadsheet etc. 46

- F -

favourite texts 23
 File Utilities: compare 2 files 148
 File Utilities: file chunker 148
 File Utilities: find duplicates 149
 File Utilities: index 145
 File Utilities: overview 7
 File Utilities: rename 149
 file-based lemmatisation 113
 file-based search-words or phrases 83
 filenames 179
 filenames display 50
 filtering 42
 finding a word 50
 finding by typing 50
 finding entries 120
 finding relevant files 39
 finding source texts 172
 first use of WordSmith 45
 folders 172
 fonts 22
 for use on pc named XXX 198
 format 40
 frequency of happi* 30

- G -

get favourite text selection 23
 getting started 3
 getting started with Concord 12
 getting started with KeyWords 13
 getting started with WordList 14
 globality of plot 180
 Greek 23
 greek font 22
 grow and shrink 77

- H -

handling multiple windows 54
 handling tag-types 58
 heading marker 52
 headings (specifying) 128
 headings: definition 170

headings: start & end 64
 hex view 158
 history list 45, 89
 holes in file 198
 horizons 69
 hotkey combinations 178
 how many words 174
 how much text 174
 how to build a database 97
 HTML
 XML 173
 HTML & SGML tags 58
 HTML headers: cutting out 60
 hyphen treatment 173
 hyphens 52

- I -

idioms 183
 illegible 190
 importing text into a word list 125
 index lists: uses 115
 information about WordSmith version 185
 installing WordSmith Tools 16
 instructions folder 16
 Internet Explorer 179
 introduction to WordSmith Tools 3
 inverted commas 188
 it won't do what I want 188

- J -

Japanese 167
 joiner 146
 joining entries 113
 joining text files 146

- K -

key key word defined 98
 key key-words 99
 key words example 97
 keyboard 178
 key-ness defined 98
 keys for searching 120
 KeyWords database 99

KeyWords: advice 99
 KeyWords: calculation 99
 KeyWords: clusters 96
 KeyWords: display 103
 KeyWords: limitations 174
 KeyWords: links 100
 KeyWords: overview 6
 KeyWords: purpose 94
 KeyWords: sorting 103
 KeyWords: starting tips 13
 KeyWords: tips 99

- L -

language 23
 Languages Chooser: font 141
 Languages Chooser: language 139
 Languages Chooser: other languages 142
 Languages Chooser: overview 138
 Languages Chooser: saving settings 142
 Languages Chooser: sort order 141
 last file 174
 layout 40
 lemma matching: WordList 108
 lemmas 113
 lemmatising with custom .dll 33
 limitations 174
 links between tools 175
 list of buttons 176
 locating entry-types 120
 log file to trace problems 36
 log likelihood 99
 Log Likelihood score 122
 logging 36
 long file names 179

- M -

machine requirements 179
 make a word list from keywords data 101
 making a tag file 62
 making Wordlist Index 116
 manual for WordSmith Tools 179
 match list 42
 memory usage 182
 menu choices 176

merge concordances 109
merge wordlists 109
MI score 122
MI3 score 122
Microsoft Word 167
Minimal Pairs 142
Minimal Pairs: aim 142
Minimal Pairs: choosing files 143
Minimal Pairs: output 143
Minimal Pairs: overview 7
Minimal Pairs: requirements 142
Minimal Pairs: rules and settings 144
Minimal Pairs: running the program 144
moving sentences 157
multimedia concordancing 76
multimedia tags 59
multiple file analysis 99
multiple lists 19
mutual information scores 120
mutual information screen 122
mutual information: computing 122

- N -

nag message 206
nearest tag 84
negative keyness 98
network defaults 48
network settings 16
network version 16
networks: defaults 48
new in version 4 5
new user 45
not a current WordSmith file 204
notes 19
number of concordance entries 90
number sort 87
numbering: paragraphs 158
numbering: sentences 158
numbers 52
numbers: how treated 180

- O -

options for defaults 48
ordering details 171

over-writing 151
Oxford University Press 171

- P -

p value 101
paragraph marker 52
paragraph numbering 158
paragraph: start & end 64
paragraphs (specifying) 128
paragraphs: definition 170
partial save 49
phrases 183
plot dispersion value 180
plot display 101
plots and links 100
plotting key words 101
popup menu 36
Portuguese 23
potato-peeling machine 190
previous lists 45
price 171
print preview 44
printer settings 181
printing 44
purple marks 77
purpose of Splitter 145
purpose of Text Converter 150
purpose of Viewer 157

- Q -

quitting 45
quotation marks 188

- R -

RAM availability 182
random deletion of entries 45
range 111, 112
reduce data to N entries 45
reference corpus 182
regrouping clumps 103
rename numerous files 149
re-ordering 54
re-ordering word lists 38

replacing 151
 research uses 67
 re-sorting a word list 129
 re-sorting: collocates 88
 re-sorting: Concord 87
 re-sorting: consistency lists 124
 re-sorting: dispersion plot 88
 re-sorting: KeyWords 103
 restore last file 174, 182
 restricted search 82
 ruler 101
 Russian 23
 russian font 22

- S -

save as HTML 46
 save as text 46
 save as XML 46
 save favourite text file set 23
 save layout 40
 save part of data 49
 saving defaults 48
 saving results 49
 search & replace 50
 search by typing 172
 search word syntax 89
 searching by typing 50
 searching for a word or part of a word 50
 searching using menu 120
 section: start & end 64
 selecting between texts 61
 selecting within texts 62
 sentence marker 52
 sentence numbering 158
 sentence: start & end 64
 sentences (specifying) 128
 sentences: definition 170
 Set column 68
 setting up a training session 40
 shortcuts 178
 show help at startup 48
 single words 183
 skimming 158
 slash 89
 slow 190
 sorting tags 84
 sorting: Concord 87
 sorting: KeyWords 103
 sorting: WordList 129
 sound & video tagged files 76
 sound file tags 59
 source texts 172
 specific limitations 174
 speed 183
 Splitter 145
 Splitter: filenames 146
 Splitter: overview 7
 Splitter: symbols 146
 Splitter: wildcards 146
 splitting 159
 standardised or mean type/token ratio 127
 start and end of sentence 64
 statistics 125
 statistics of a database 99
 status bar 176, 184
 stop lists 51
 stoplist.cod 154
 stopping 51
 storage 182
 store text files 24
 student use 67
 summary statistics 30
 suspending processing 51
 symbols 166

- T -

tag concordancing 81
 tag context 79
 tag file 62
 tag types 57
 tagged text 57
 tags as selectors 60
 tags: overview 57
 teacher instructions 40
 teaching uses 67
 text characteristics 52
 Text Converter folders 151
 Text Converter settings 151
 Text Converter: asterisk 152
 Text Converter: conversion file 154

Text Converter: limitations 174
Text Converter: move if 154
Text Converter: overview 7
Text Converter: sample conversion file 155
Text Converter: syntax 152
Text Converter: wildcards 152
text file: use to build a word list 125
text formats 52
text segments in Concord 88
texts: choosing 24
texts: more texts 24
the ~ operator 82
tie-breaking 87
too many requests to ignore matching clumps 206
too many sentences 206
toolbar 176
tools for pattern-spotting 184
training students 40
troubleshooting 188
troubleshooting: accented symbols 189
troubleshooting: apostrophes not found 188
troubleshooting: colours unreadable 190
troubleshooting: column spacing 188
troubleshooting: Concord tags problem 188
troubleshooting: Concord/WordList mismatch 189
troubleshooting: crashed 189
troubleshooting: curly quotation marks 188
troubleshooting: demo limit 189
troubleshooting: keys don't respond 190
troubleshooting: pineapple-slicing 190
troubleshooting: printer won't print 190
troubleshooting: quotation marks not found 188
troubleshooting: smart quotations 188
troubleshooting: takes ages 190
troubleshooting: Viewer 160
troubleshooting: weird symbols 189
troubleshooting: won't start 191
troubleshooting: WordList out of order 191
truncating at xx words 206
two files needed 207
Two word-list analysis 94
type/token ratios 127
typeface 40
type-in mode 172
type-in search 50
types of tag 57

- U -

Unicode codes 167
university or school work 40
unjoining 113
unreadable 190
updater.exe 171
updating your version 171
user-defined categories 68
user-defined process 33

- V -

version 4 differences 163
Version Checker: overview 8
version checking 17
version date 185
version mis-match 208
Viewer 157
Viewer: aligning the sentences 157
Viewer: editing 158
Viewer: highlighted words 158
Viewer: limitations 174
Viewer: overview 9
Viewer: sentence joining 159
Viewer: settings 159
Viewer: skimming 158
Viewer: tags 158
Viewer: technical aspects 159
Viewer: translation mis-matches 160
Viewer: unusual sentences 160
Viewer: viewing options 158
viewing original text file 77

- W -

WebGetter: display 137
WebGetter: limitations 138
WebGetter: overview 135, 9
WebGetter: settings 135
what is a concordance 67
What's new 5
whole word search 89
why did search fail? 207
why won't it... 188

window management 54
Windows 2000 179
Windows 95 filenames 179
Windows 98 179
Windows character set codes 167
Windows NT 179
Windows XP 179
word list file not found 207
word list is faulty 207
word patterns 86
word separators 171
word: definition 170
WordList comparison file faulty 207
WordList index lists: viewing 115
WordList overview 6
WordList: altering entries 38
WordList: case sensitivity 128
WordList: clusters 117
WordList: create using text file 125
WordList: limitations 174
WordList: minimum & maximum settings 128
WordList: purpose 107
WordList: sort 191
WordList: sort order 129
WordList: starting tips 14
WordList: the basic display 130
WordSmith already running 207
WordSmith controller: Concord: settings 90
WordSmith controller: KeyWords settings 104
WordSmith controller: WordList settings 132
WordSmith Tools: installation 16
WordSmith Tools: manual 179
WordSmith version 185
wshell.exe (controller) 5
wshell.ini and networks 16

- X -

XX days left 208

- Z -

Z score 122
zapping 54
zip files 186

Endnotes 2... (after index)

Back Cover