WordSmith Tools Manual

Version 7.0

© 2019 Mike Scott

Lexical Analysis Software Ltd.
Stroud, Gloucestershire, UK

WordSmith Tools Manual

version 7.0

---

by Mike Scott

Compiled: October, 2019

WordSmith Tools Manual

© 2019 Mike Scott

All rights reserved. But most parts of this work may be reproduced in any form or by any means - graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems - usually without the written permission of the publisher.
See http://www.lexically.net/publications/copyright_permission_for_screenshots.htm

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

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

Produced: October 2019

Publisher

Lexical Analysis Software Ltd.

Special thanks to:

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

Table of Contents

Foreword 0

Part I WordSmith Tools 1

Part II Overview 3

1 What's new in this version ... 4
2 Controller ... 5
3 Concord ... 6
4 KeyWords ... 7
5 WordList ... 7
6 Utilities ... 8
Aligner ... 8
Character Profiler ... 9
CharGrams ... 9
Choose Languages ... 9
Corpus Checker ... 10
File Utilities ... 10
Splitter ... 10
File Viewer ... 10
Minimal Pairs ... 11
Text Converter ... 11
Version Checker ... 11
WSConcGram ... 13

Part III Getting Started 14

1 getting started with Concord ... 15
2 getting started with KeyWords ... 16
3 getting started with WordList ... 18

Part IV Installation and Updating 21

1 installing WordSmith Tools ... 22
2 what your licence allows ... 24
3 site licence defaults ... 25
4 version checking ... 25

Part V Controller 27

1 add notes ... 28
2 batch processing ... 30
3 category or Set ... 33
add a Set column ... 33
colour categories ... 35
set column categories ... 43

© 2019 Mike Scott

colour categories ... 45
4 choosing texts ... 49
text formats ... 49
the file-choose window ... 52
favourite texts ... 59
5 class or session instructions ... 59
6 columns of data ... 60
layout & format ... 60
compute new column of data ... 65
column totals ... 67
7 compute concordance ... 69
8 copy your results ... 69
9 editing ... 70
reduce data to n entries ... 71
reverse-delete ... 71
delete or restore to end ... 71
delete if ... 71
editing column headings ... 74
editing a list of data ... 75
10 language ... 77
Change language within a Tool ... 79
Language Chooser ... 79
Overview ... 80
saving your choices ... 81
Font ... 81
Other Languages ... 82
11 match words in list ... 82
12 plot contexts ... 86
13 previous lists ... 87
14 print and print preview ... 88
15 quit WordSmith ... 91
16 saving ... 91
save results ... 91
save as text, Excel etc. ... 91
17 searching ... 95
three alternatives ... 95
Find ... 96
Colour categories find ... 99
Search text ... 101
Search & replace ... 102
find relevant files ... 104
search by typing ... 106
18 selecting or marking entries ... 107
selecting multiple entries ... 107
marking entries ... 108
19 settings ... 108
restoring settings ... 108
save defaults ... 110

© 2019 Mike Scott

adjust settings ... 111
main settings ... 113
colours ... 114
advanced colour settings ... 117
folder settings ... 118
fonts ... 118
advanced settings ... 120
scripting ... 123
custom processing ... 128
percentage columns ... 131
Help_and_logging ... 131
clipboard ... 134
20 source text(s) ... 134
21 breakdowns ... 138
count data frequencies ... 141
22 stop lists ... 141
23 suspend processing ... 144
24 text and languages ... 146
characters and letters ... 147
accents and other characters ... 147
wildcards ... 148
25 text dates and time-lines ... 149
26 view ... 150
27 window management ... 151
filenames tab ... 152
28 word clouds ... 152
29 zap unwanted lines ... 154
Part VI Tags and Markup 156
1 choices in handling tags ... 158
2 tag-types ... 159
multimedia tags ... 162
start and end of text segments ... 164
3 custom settings ... 165
4 tags as selectors ... 167
tags to select between texts ... 168
part of file:selecting within texts ... 168
only if containing... ... 171
5 making a tag file ... 172
6 modify source texts ... 182
7 examples ... 186
only words with tag-file tags ... 187
tags but no words ... 188
8 XML text such as BNC XML ... 189
Part VII Concord 194
1 index ... 195

© 2019 Mike Scott

2 what is a concordance? ... 196
3 concordance view ... 196
4 what can I do with the list? ... 198
blanking ... 199
source text context ... 199
5 breakdowns ... 200
6 clusters ... 203
7 Collocation ... 206
what is collocation? ... 206
collocate horizons ... 207
collocation relationship ... 207
collocates and relative entropy ... 209
collocates display ... 210
collocates and lemmas ... 211
collocate follow ... 213
collocate highlighting in concordance ... 215
follow-up concordances ... 216
collocate settings ... 217
re-sorting: collocates ... 219
8 concordancing on tags ... 221
nearest tag ... 223
9 context word ... 226
10 dispersion plot ... 230
11 editing concordances ... 236
remove duplicates ... 236
12 patterns ... 237
13 re-sorting ... 238
re-sorting: dispersion plot ... 240
14 file-names data ... 240
recompute file lists ... 242
15 saving and printing ... 242
16 search-word or phrase ... 246
search word syntax ... 246
file-based search-words ... 247
search-word and other settings ... 249
copy search-word to Set column ... 252
17 settings ... 253
what you GET ... 253
what you SEE ... 257
18 sounds & video ... 258
playing the video ... 261
obtaining sound and video files ... 263
19 text segments in Concord ... 263

Part VIII KeyWords 265

1 index ... 266

© 2019 Mike Scott

2 ordinary two word-list analysis ... 267
3 choosing files ... 269
4 example of key words ... 272
5 keyness ... 273
p value ... 273
keyness definition ... 273
thinking about keyness ... 274
6 KeyWords database ... 275
creating a database ... 276
key key-word definition ... 278
associates ... 279
associate definition ... 281
follow-up ... 282
keywords database related clusters ... 285
export data ... 285
7 KeyWords: advice ... 286
8 KeyWords: calculation ... 287
9 KeyWords: text dispersion ... 288
10 KeyWords clusters ... 291
11 KeyWords: links ... 292
12 make a word list from keywords data ... 293
13 plot display ... 294
plot calculation ... 296
14 re-sorting: KeyWords ... 296
15 settings ... 297
what you GET ... 297
what you SEE ... 300
database settings ... 300
16 the key words screen ... 302

Part IX WordList 304

1 index ... 305
2 clusters ... 306
3 compare word lists ... 307
compute key words ... 307
comparing wordlists ... 308
comparison display ... 309
4 merging wordlists ... 310
5 consistency ... 310
consistency analysis (range) ... 310
detailed consistency analysis ... 311
re-sorting: consistency lists ... 319
detailed consistency relations ... 319
6 find filenames ... 320
7 Lemmas (joining words) ... 321
how do we join words? ... 322

© 2019 Mike Scott

manual joining ... 325
auto-joining lemmas ... 326
choosing lemma file ... 327
8 WordList Index ... 330
what is an Index for? ... 330
making a WordList Index ... 330
index clusters ... 332
join clusters ... 337
index lists: viewing ... 338
index exporting ... 340
9 menu search ... 341
10 relationships between words ... 342
mutual information and other relations ... 342
relationships display ... 342
relationships computing ... 344
11 recompute tokens ... 349
12 statistics ... 350
type/token ratios ... 353
dispersion ... 354
relative entropy ... 356
summary statistics, breakdowns ... 359
13 stop-lists and match-lists ... 362
14 import words from text list ... 363
15 settings ... 365
WordSmith controller: Index settings ... 366
WordSmith Controller WordList settings ... 367
what you GET ... 367
what you SEE ... 369
minimum & maximum settings ... 370
case sensitivity ... 371
16 sorting ... 372
17 WordList and tags ... 374
18 WordList display ... 380

Part X Utility Programs 385

1 Aligner ... 387
index ... 387
aligning ... 388
example of aligning ... 389
TED talks ... 391
aligning and moving ... 392
editing ... 393
finding specific words ... 393
languages ... 395
numbering sentences & paragraphs ... 395
options ... 395
parallel concordancing ... 396
reading in a plain text ... 398
joining and splitting ... 400

© 2019 Mike Scott

settings ... 401
technical aspects ... 401
translation mis-matches ... 402
troubleshooting ... 402
unusual lines ... 403
2 Character Profiler ... 404
purpose ... 404
profiling characters ... 405
converting ... 407
profiling settings ... 410
3 Chargrams ... 411
purpose ... 411
chargram procedure ... 411
display ... 414
patterns ... 416
settings ... 418
4 Corpus Checker ... 419
Aim ... 419
checking corruption ... 420
find duplicates ... 422
similar content ... 422
same names ... 425
boilerplate text ... 426
relevance check ... 429
Filter relevant texts ... 434
corpus sampler ... 434
RTF copy ... 436
build sub-corpora ... 438
5 File Utilities ... 439
index ... 439
compare two files ... 440
dates and times ... 440
highlighter ... 442
lemma file building ... 444
tags ... 446
move files to sub-folders ... 448
count and/or delete ... 450
rename ... 451
Splitter ... 453
Splitter: index ... 453
Splitter settings ... 454
Splitter: filenames ... 456
Splitter: wildcards ... 457
join text files ... 457
file chunker ... 458
find holes in texts ... 459
6 File Viewer ... 459
Using File Viewer ... 459
7 Minimal Pairs ... 462
aim ... 462
requirements ... 462
choosing your files ... 464

© 2019 Mike Scott

output ... 466
rules and settings ... 467
running the program ... 468
8 Registration ... 471
9 Text Converter ... 471
index ... 472
settings ... 473
Text Converter: copy to ... 477
extracting from files ... 478
filtering: move if ... 479
Convert within the text file ... 481
conversion file ... 484
syntax ... 485
sample conversion file ... 486
remove or replace ... 487
multi-word units ... 488
Convert format of entire text files ... 489
Mark-up changes ... 492
Word, Excel, PDF, RTF, TMX ... 497
non-Unicode Text ... 500
Other changes ... 501
Text Converter: converting BNC XML version ... 504
10 WSConcGram ... 510
aims ... 510
definition of a concgram ... 511
settings ... 512
generating concgrams ... 513
viewing concgrams ... 514
filtering concgrams ... 519
exporting concgrams ... 520
Part XI Reference 522
1 acknowledgements ... 523
2 API ... 523
3 bibliography ... 524
4 bugs ... 525
5 change language ... 526
6 Character Sets ... 527
overview ... 527
accents & symbols ... 528
7 clipboard ... 529
8 contact addresses ... 532
9 date format ... 532
10 Definitions ... 533
definitions ... 533
word separators ... 534
11 demonstration version ... 534
12 drag and drop ... 534

© 2019 Mike Scott

13 edit v. type-in mode ... 535
14 file extensions ... 536
15 finding source texts ... 537
16 flavours of Unicode ... 538
17 folders\directories ... 539
18 formulae ... 540
19 history list ... 543
20 HTML, SGML and XML ... 543
21 hyphens ... 544
22 international versions ... 544
23 limitations ... 544
tool-specific limitations ... 544
24 links between tools ... 545
25 keyboard shortcuts ... 546
26 machine requirements ... 548
27 manual for WordSmith Tools ... 548
28 menu and button options ... 548
29 MS Word documents ... 552
30 never used WordSmith before ... 554
31 numbers ... 554
32 plot dispersion value ... 554
33 RAM availability ... 555
34 reference corpus ... 555
35 restore last file ... 555
36 single words v. clusters ... 556
37 speed ... 557
38 status bar ... 558
39 tools for pattern-spotting ... 558
40 version information ... 559
Version 3 improvements ... 560
41 zip files ... 561

Part XII Troubleshooting 562

1 list of FAQs ... 563
2 apostrophes not found ... 563
3 column spacing ... 563
4 Concord tags problem ... 563
5 Concord/WordList mismatch ... 564
6 crashed ... 564
7 demo limit ... 564

© 2019 Mike Scott

8 funny symbols ... 564
9 illegible colours ... 565
10 keys don't respond ... 565
11 pineapple-slicing ... 566
12 printer didn't print ... 566
13 too slow ... 566
14 won't start ... 566
15 word list out of order ... 567

Part XIII Error Messages 568

1 list of error messages ... 569
2 .ini file not found ... 570
3 administrator rights ... 570
4 base list error ... 571
5 can only save words as ASCII ... 571
6 can't call other tool ... 571
7 can't make folder as that's an existing filename ... 571
8 can't compute key words as languages differ ... 571
9 can't merge list with itself! ... 571
10 can't read file ... 571
11 character set reset to <x> to suit <language> ... 572
12 concordance file is faulty ... 572
13 concordance stop list file not found ... 572
14 confirmation messages: okay to re-read ... 572
15 conversion file not found ... 572
16 couldn't save file in folder ... 573
17 destination folder not found ... 573
18 disk problem -- file not saved ... 573
19 dispersions go with concordances ... 573
20 drive not valid ... 573
21 failed to access Internet ... 573
22 failed to create new folder name ... 573
23 failed to read file ... 573
24 failed to save file ... 574
25 file access denied ... 574
26 file contains none of the tags specified ... 574
27 file has "holes" ... 574
28 file not found ... 574
29 filenames must differ! ... 575
30 folder is read-only ... 575

© 2019 Mike Scott

31 for use on X machine only ... 575
32 form incomplete ... 575
33 full drive & folder name needed ... 575
34 function not working properly yet ... 575
35 invalid concordance file ... 575
36 invalid file name ... 576
37 invalid KeyWords database file ... 576
38 invalid KeyWords calculation ... 576
39 invalid WordList comparison file ... 576
40 invalid WordList file ... 576
41 joining limit reached ... 577
42 KeyWords database file is faulty ... 577
43 KeyWords file is faulty ... 577
44 limit of file-based search-words reached ... 577
45 links between Tools disrupted ... 577
46 match list details not specified ... 577
47 must be a number ... 578
48 mutual information incompatible ... 578
49 network registration used elsewhere ... 578
50 no access to text file - in use elsewhere? ... 578
51 no associates found ... 578
52 no clumps identified ... 578
53 no clusters found ... 578
54 no collocates found ... 579
55 no concordance entries ... 579
56 no concordance stop list words ... 579
57 no deleted lines to zap ... 579
58 no entries in KeyWords database ... 579
59 no fonts available ... 579
60 no key words found ... 580
61 no key words to plot ... 580
62 no KeyWords stop list words ... 580
63 no lemma list words ... 580
64 no match list words ... 580
65 no room for computed variable ... 580
66 no statistics available ... 580
67 no stop list words ... 580
68 no such file(s) found ... 581
69 no tag list words ... 581

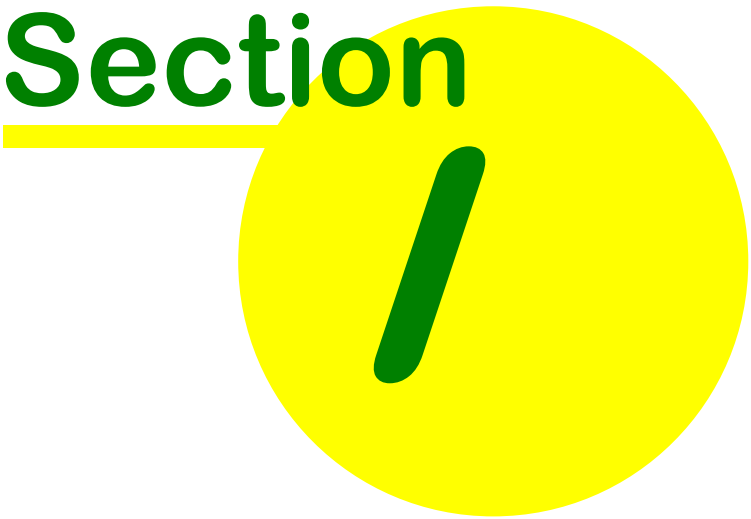
© 2019 Mike Scott

70 no word lists selected ... 581
71 not a valid number ... 581
72 not a WordSmith file ... 581
73 not a current WordSmith file ... 581
74 nothing activated ... 582
75 Only X% of words found in reference corpus ... 582
76 original text file needed but not found ... 582
77 printer needed ... 582
78 registration code in wrong format ... 583
79 registration is not correct ... 583
80 short of memory ... 583
81 source folder file(s) not found ... 583
82 stop list file not found ... 583
83 stop list file not read ... 583
84 tag file not found ... 583
85 tag file not read ... 584
86 this function is not yet ready ... 584
87 this is a demo version ... 584
88 to stop getting this message 584
89 too many requests to ignore matching clumps ... 584
90 too many sentences ... 584
91 truncating at xx words -- tag list file has more ... 584
92 two files needed ... 584
93 unable to merge Keywords Databases ... 584
94 why did my search fail? ... 585
95 word list file is faulty ... 585
96 word list file not found ... 585
97 WordList comparison file is faulty ... 585
98 WordSmith Tools already running ... 585
99 WordSmith Tools expired ... 585
100 WordSmith version mis-match ... 585
101 XX days left ... 586

Index 587

© 2019 Mike Scott

WordSmith Tools



1 WordSmith Tools

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

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

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

Getting Help

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

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

See also: getting started straight away with WordList 18, Concord 15, or KeyWords 16.

Version:7.0

© 2019 Mike Scott

Overview

2 Overview

WordSmith Tools requires

1. a reasonably up-to-date computer 548
2. running Windows XP 548 or later
3. your own collection of text in plain text format, HTML or XML format, or converted 489 to plain text

2.1 What's new in this version

WordSmith is organic software!

Version 1.0 was launched in 1996. We have continued an organic policy of growth ever since ... now in 2019 we are at version 7.0 with improvements and new features.

New features:

- Multi-word unit conversion 482
- Plot contexts 86
- Concord files processed list (automatic 122 or manual 240)
- Add a Set column 33
- Export KeyWords database data 285
- TED talks video conversion 502
- Enhanced key word statistics 288 and key word display 302
- Part of speech tag filtering 377
- Multiple statistics 207 in collocation display 210
- Dispersion 354 statistic for word lists
- Lemma file building 444
- Mis-matched tag 446 finder
- Boilerplate text 426 finding
- Duplicate and near-duplicate text file 422 finding
- Relevance filter 429
- Highlighter 442
- BIC scores 288 for key items
- Corpus sampler 434
- choice for percentage columns 131
- improved colours 114
- high definition screen focus
- Text Converter case change 504
- Text Converter remove words 487
- Relative entropy 356
- Associates follow-up 282
- Chargram patterns 416
- Part of speech renaming 178

- Conditional Probability statistics 207
- Build sub-corpora 438
- Clusters in single-word word list 306
- TMX file handling 499
- Copy search-word to Set column 252
- File count facility 450
- RTF Sample 436
- Normalised counts 311
- Text dispersion keyness 288

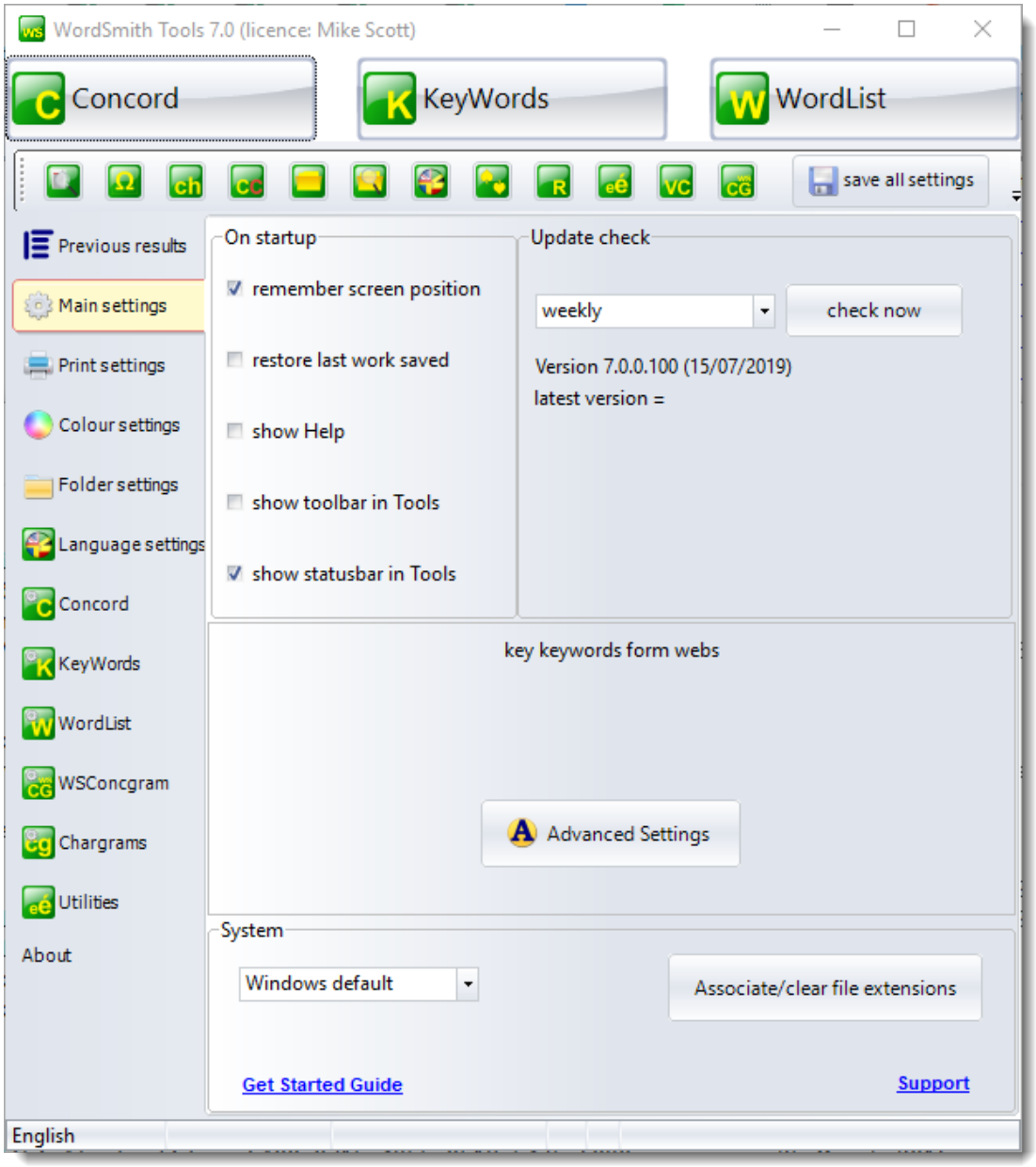
See also : new in Version 6, new in Version 5

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 151.



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 196 using plain text or web text files.

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

Listings can be saved 242 for later use, edited, printed, copied to your word-processor, or saved as text files.

See also: Concord Help Contents Page 195

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 296.

Listings can be saved 91 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 7 program.

See also: KeyWords Help Contents Page 266

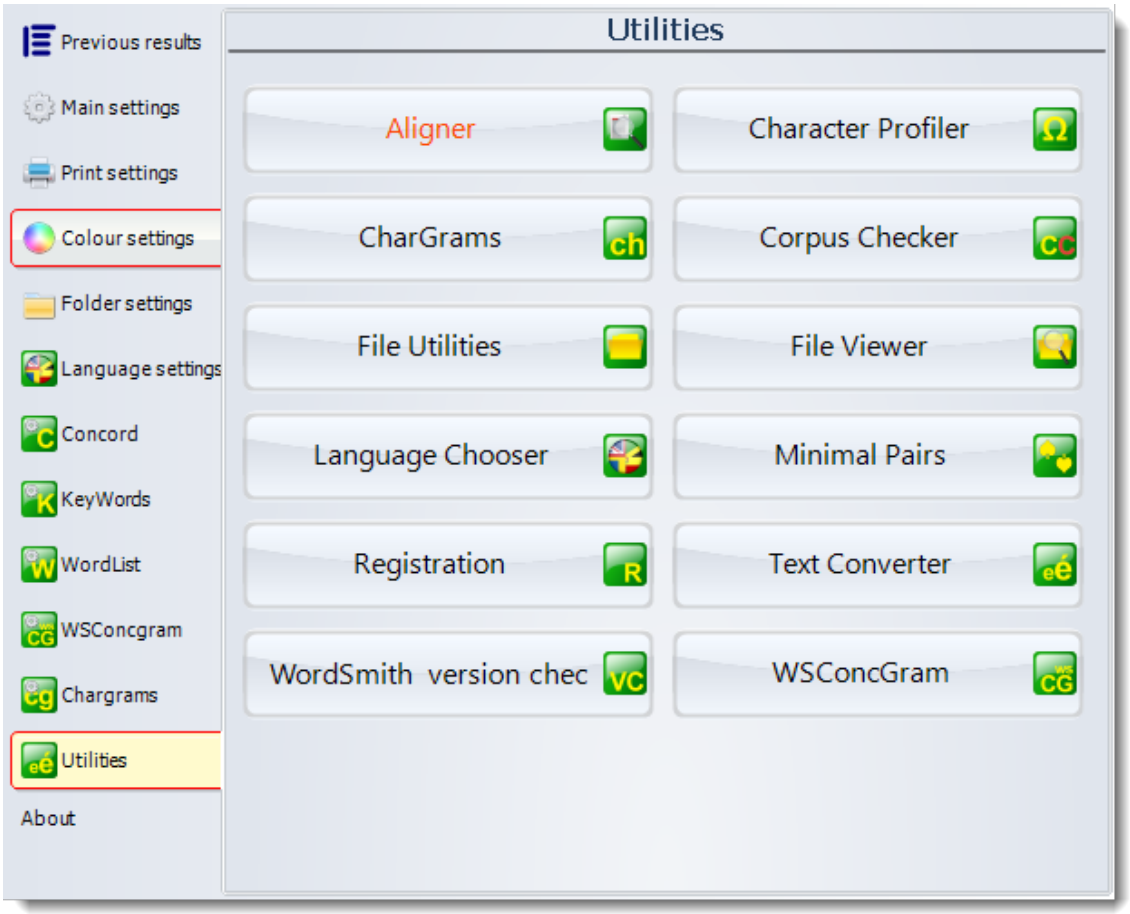
2.5 WordList

WordList generates word lists based on one or more plain text 527 or web text files. Word lists are shown both in alphabetical and frequency order. They can be saved 91 for later use, edited, printed, copied to your word-processor, or saved as text files.

See also: WordList Help Contents Page 305

2.6 Utilities

The program suite also contains a number of other programs for specific jobs. Start them using these buttons:



See also : details on each utility 386

2.6.1 Aligner

Aligner is a tool which enables you to produce a copy of a text file with numbered sentences or paragraphs 395 and then align 388 two or more versions of a text, showing alternate paragraphs or sentences of each.

See also: Aligner Help Contents Page 387

2.6.2 Character Profiler

A tool to help find out which characters or chargrams 534 are most frequent in a text or a set of texts.

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

See also: Character Profiling 404

2.6.3 CharGrams

A tool to help find out which chargrams 534 (sequences of characters) are most frequent in a text or a set of texts.

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

See also: Chargrams Tool 411

2.6.4 Choose Languages

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

See also: Choose Language Tool 79

2.6.5 Corpus Checker

A tool to go through your corpus and seek out any text files which may have become corrupted. Works in any language. Also, tests for boilerplate text, file-duplication, relevance

See also: detecting corpus corruption 420

2.6.6 File Utilities

Utilities to

- compare two files 440
- cut large files into chunks 458
- find duplicate files 425
- rename 451 multiple files
- split large files into their component texts 453
- join up 457 a lot of small text files into merged text files
- find holes 459 in your text files

2.6.6.1 Splitter

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

See also: Splitter Help Contents Page 453

2.6.7 File Viewer

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

See also: File Viewer Index 79

2.6.8 Minimal Pairs

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

See also : aim 462, requirements 462, choosing your files 464, output 466, rules and settings 467, running the program 468.

2.6.9 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.

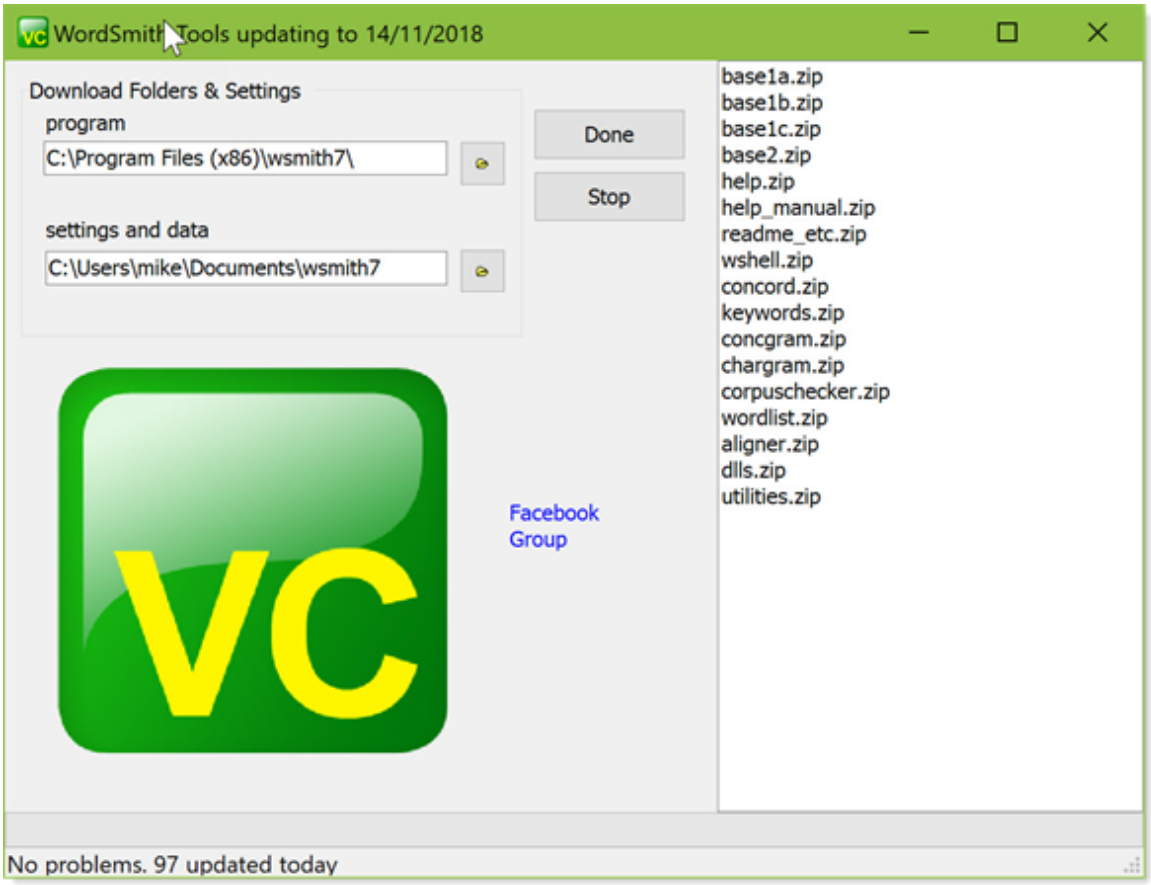
It is very useful for going through large numbers of texts and re-formatting them as you prefer, e.g. taking out unnecessary spaces, ensuring only paragraphs have <Enter> at their ends, changing accented characters.

See also: Text Converter Help Contents Page 472

2.6.10 Version Checker

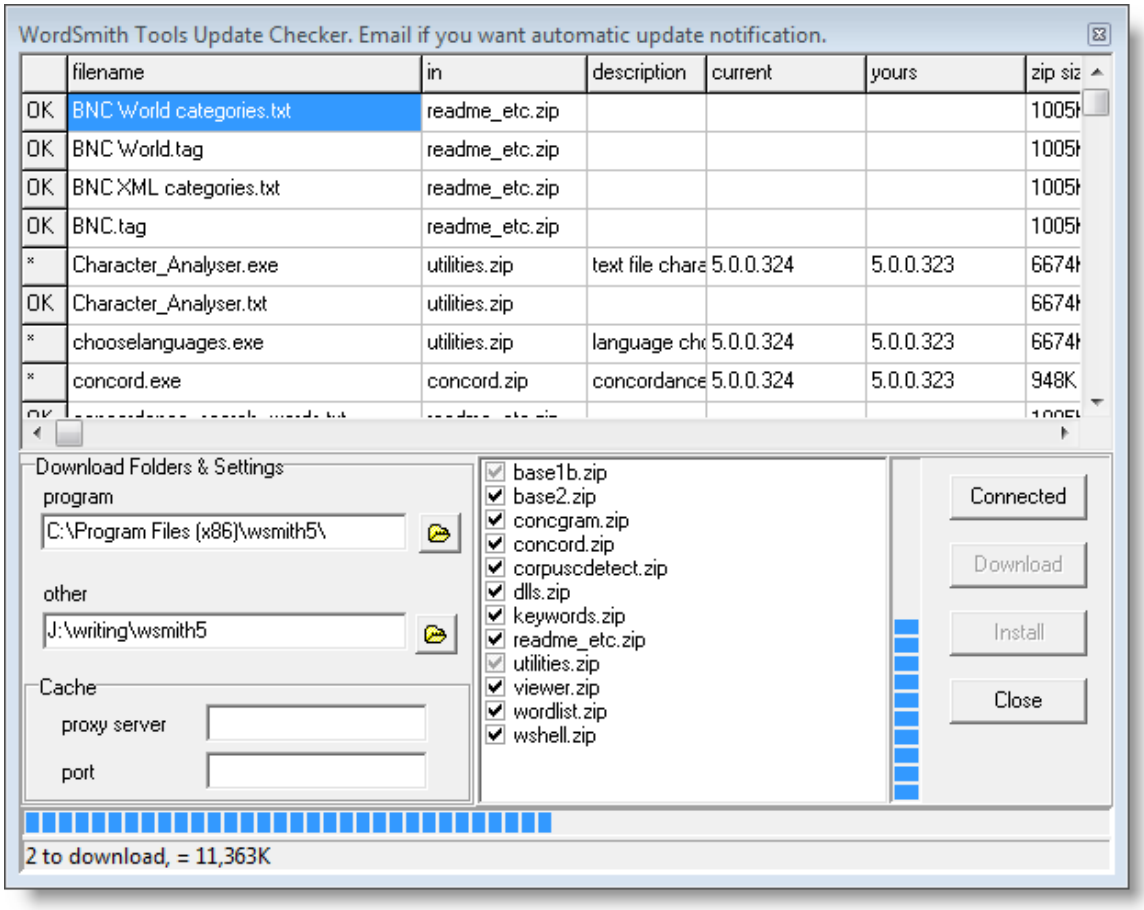
A tool to check whether any components of your current version need updating and if so, download them for you. Accessed via the main Controller menu, *File* | *Web version check*.

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



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

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



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.

2.6.11 WSConcGram

a tool for generating concgrams[511].

See also : Aims of WSConcGram[510], Running WSConcGram[513]

Getting Started

3 Getting Started

For a step-by-step view with screen-shots in a number of languages, visit the WordSmith website.

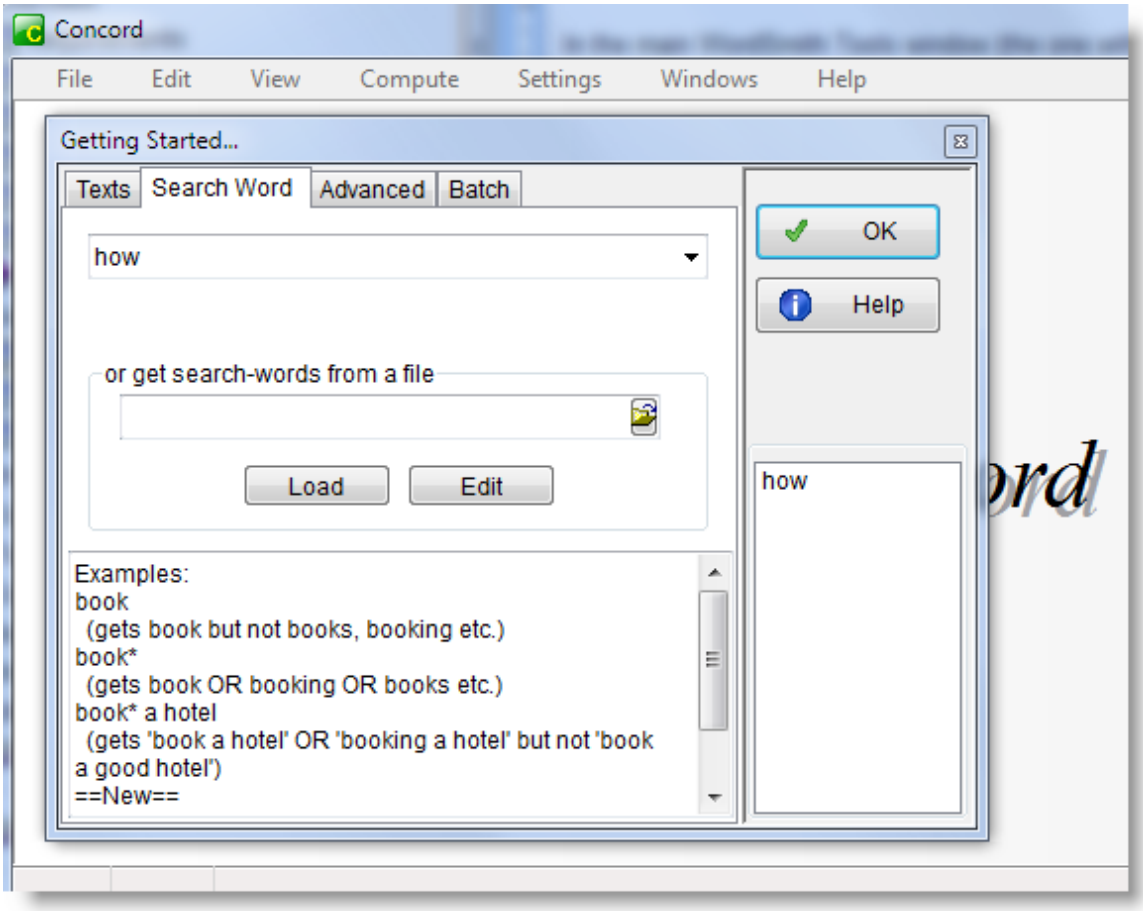
This chapter gives you advice on quickly starting up the Concord 15, KeyWords 16 and WordList 18 tools.

3.1 getting started with Concord

In the main WordSmith Tools window (the one with WordSmith Tools Controller 5 in its title bar), choose the Tools option, and once that's opened up, you'll see the Concord button. Click and the Concord tool will start up.

Choose *File* | *New*

You should now see a dialogue box which lets you choose your texts 52 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 554 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 246 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 253, 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 221.

Don't forget to save the results 242 (press Ctrl+F2 or) if you want to keep the concordance for another time.

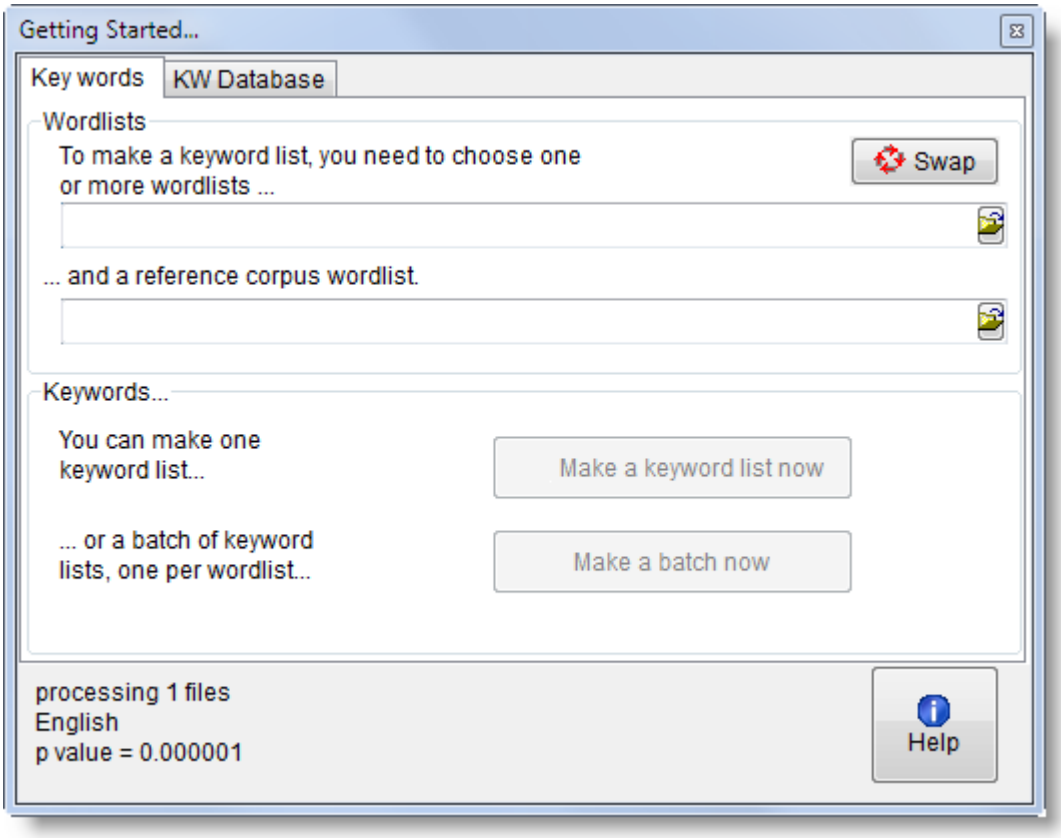
See also: Concord Help Contents 195.

3.2 getting started with KeyWords

In the main WordSmith Tools window (the one with WordSmith Tools Controller 5 in its title bar), choose the Tools option, and once that's opened up, you'll see KeyWords. Click and KeyWords will open up.

Choose *File* | *New*

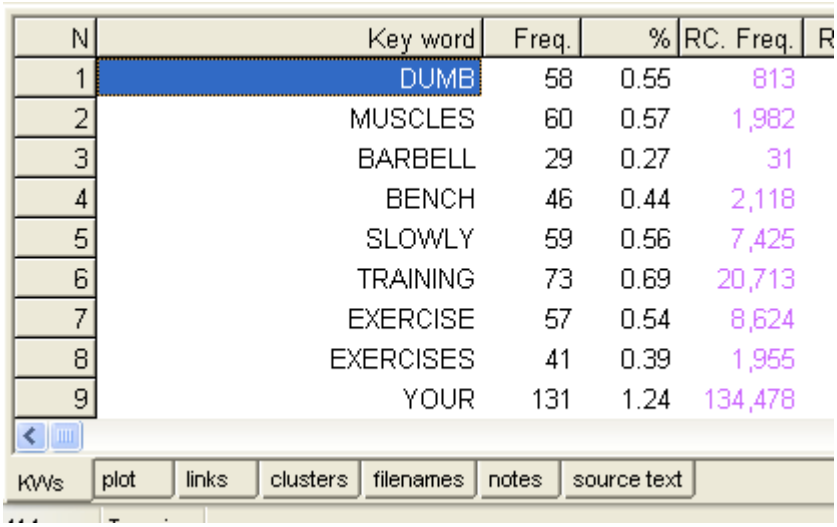
You see a dialogue box which lets you choose your word-lists 269.



You'll need to choose two word lists to make a key words list from: one based on a single text (or single corpus), and another one based on a comparison corpus of texts or reference corpus 555.

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 comparison corpus 555 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 144 window in the main Controller shows a bar indicating how it's getting on. After KeyWords has finished, it will show you a list of the key words. The ones at the top are more "key" than those further down.



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

Don't forget to save the results 91 (press Ctrl+F2 or) if you want to keep the keyword list for another time.

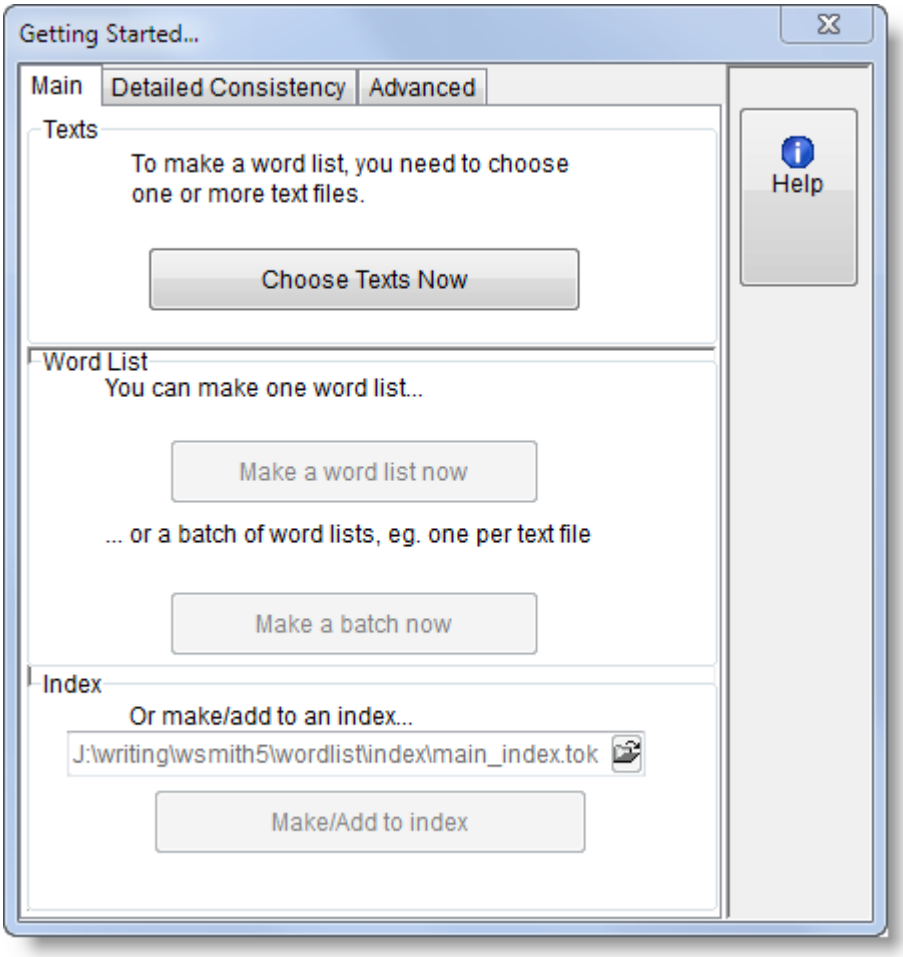
See also: KeyWords Help Contents 266, What's it for? 266

3.3 getting started with WordList

I suggest you start by trying the WordList program. In the main WordSmith Tools window (the one with WordSmith Tools Controller 5 in its title bar), choose the Tools option, and once that's opened up, you'll see WordList. Click and WordList will open up.

Choose *File* | *New*

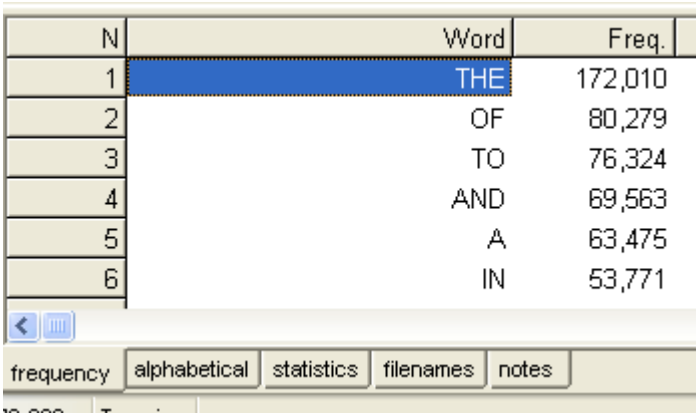
You will see a dialogue box which lets you choose your texts 52 or change your choice, and make a new word list.



If you have never used WordSmith 554 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 30.

You'll find that WordList starts processing your file(s) and a progress 144 window in the main Controller shows a bar indicating how it's getting on. After WordList has finished making the list, you will see some windows showing the words from your text file in alphabetical order and in frequency order, statistics, filenames, notes 28.



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

frequency | alphabetical | statistics | filenames | notes

Don't forget to save the results 91 (press Ctrl+F2 or) if you want to keep the word list for another time.

See also: WordList Help Contents 305.

Installation and Updating

Section

4 Installation and Updating

Download the program from the downloads page.

The setup program will expand all the files needed for WordSmith Tools into the folder of your choice (**\wsmith7** by default). You can install to a removable drive if you wish (explained below).

Now run **\wsmith7\wordsmith7.exe** to get started. You will be asked to register. Otherwise WordSmith will go through its paces as a Demonstration Version 534.

Install on a USB stick?

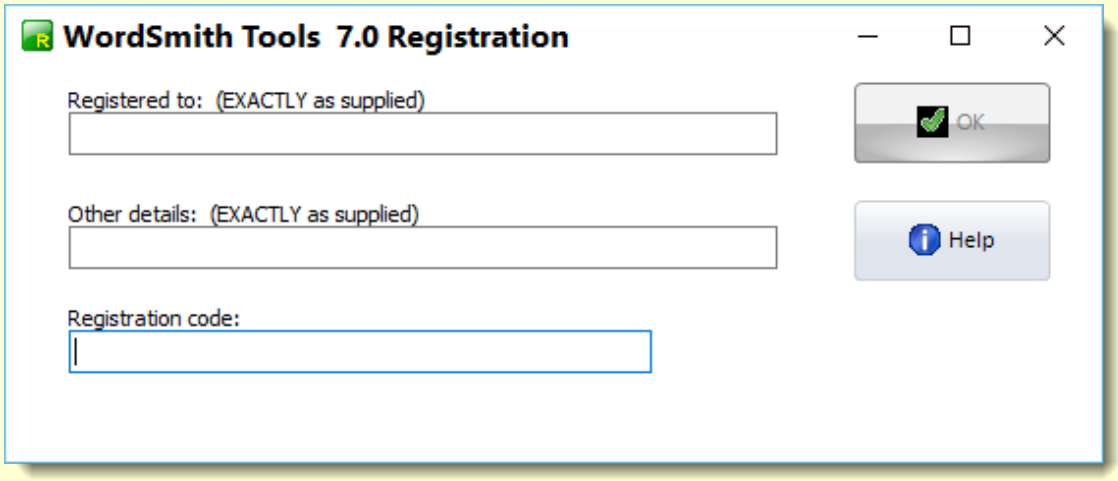
WordSmith will run happily from a memory stick or external drive. Many users find it convenient to install on an external USB drive or memory stick. That way you can easily take it with you and run on a lot of different PCs. Many universities do let software run from a USB drive even though they prevent their users from installing. More details here 23.

4.1 installing WordSmith Tools

Run **\wsmith7\wordsmith7.exe** to get started. You will be asked to register. Otherwise WordSmith will go through its paces as a Demonstration Version 534.

Upon receipt of the registration code, run WordSmith Tools. If you have only just installed, the registration program 471 will start up automatically. If not you can run **\wsmith7\WSRegister7.exe**. or find it in the Utilities 386 section of the Controller.

Single User Licence



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

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

This name appears in the main window and whenever you access the About option (F9). Your software will then be fully enabled, and the Update from Demo menu option will disappear. (The **WSRegister7.exe** program will still be there in your \wsmith7 folder, and can be used if you ever need to re-register.)

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

install to a removable drive

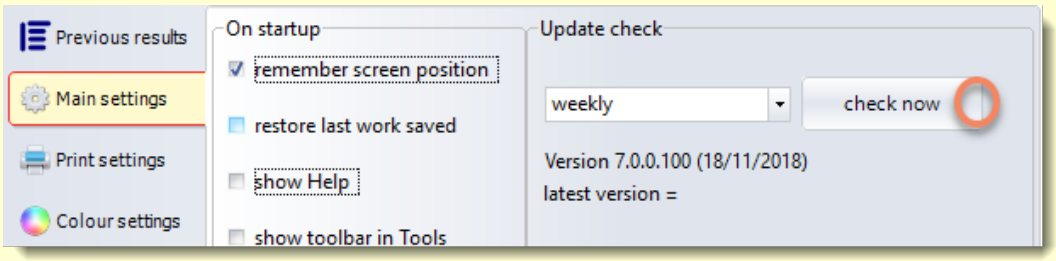
You don't need to install to the **C:** drive -- you can install WordSmith on a USB drive such as a pen drive or memory stick, or a fast external hard drive. That way you can take WordSmith with you from one computer to the next. A cheap USB drive will be a rather slow medium, but a fast external drive can be very satisfactory in terms of speed. If you save your default settings 110, any folder names 539 which are on the external drive itself get the drive letter corrected automatically as do folder names in any data you save.

Site Licence

Follow the instructions at
https://lexically.net/wordsmith/version7/faqs/network_installation.htm.

Updating your version

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



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

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

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

To un-install, just delete all the files in your **\wsmith7** folder. Your data may be in sub-folders of **\wsmith7** or in sub-folders of your **Documents\wsmith7.**

See also: Setting default options 110, Contact Addresses 532, File types 536.

4.2 what your licence allows

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

SINGLE USER LICENCES

Think of these as a licence for a person.

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

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

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

SITE LICENCES

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

The full licence text is at **\wsmith7\user_licence.txt**. Your licence doesn't expire but it is limited to the major version you bought. So a WordSmith 5 licence won't work with WordSmith 6 and a WS 6 licence won't work with WS7 but you can re-download the program as needed from our site.

4.3 site licence defaults

If you have bought a site licence, just install one copy of WordSmith on any shared drive accessible by all your users. Follow the instructions at

https://lexically.net/wordsmith/version7/faqs/network_installation.htm.

the wordsmith7.ini file

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

example

```
[NETWORK]
network-read/write folder=m:\Documents\wsmith7
```

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

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

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

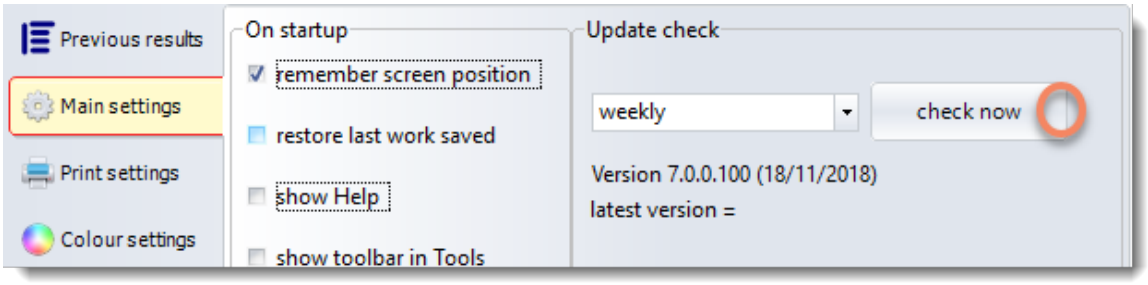
When a new user starts using WordSmith for the very first time, WordSmith will notice that it is running on a site-licence version and read the "network-read/write folder" information above. It will then try to automatically create the folder you have specified above (in theory you shouldn't need to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

For networked drives, because of a Microsoft security update involving HTML Help files, WordSmith will copy the **wordsmith7.chm** file to the user's Windows-allocated temporary folder.

See also: What your licence allows 24, Class Instructions 59

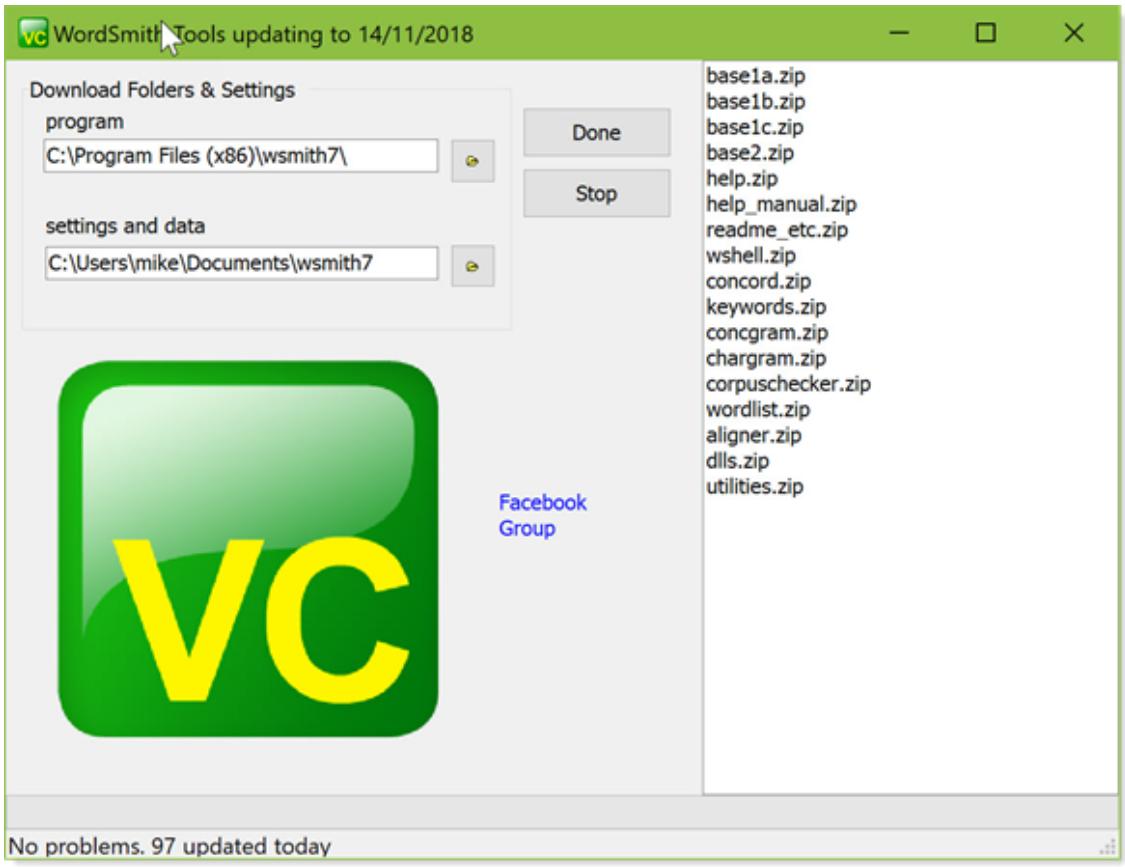
4.4 version checking

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



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

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



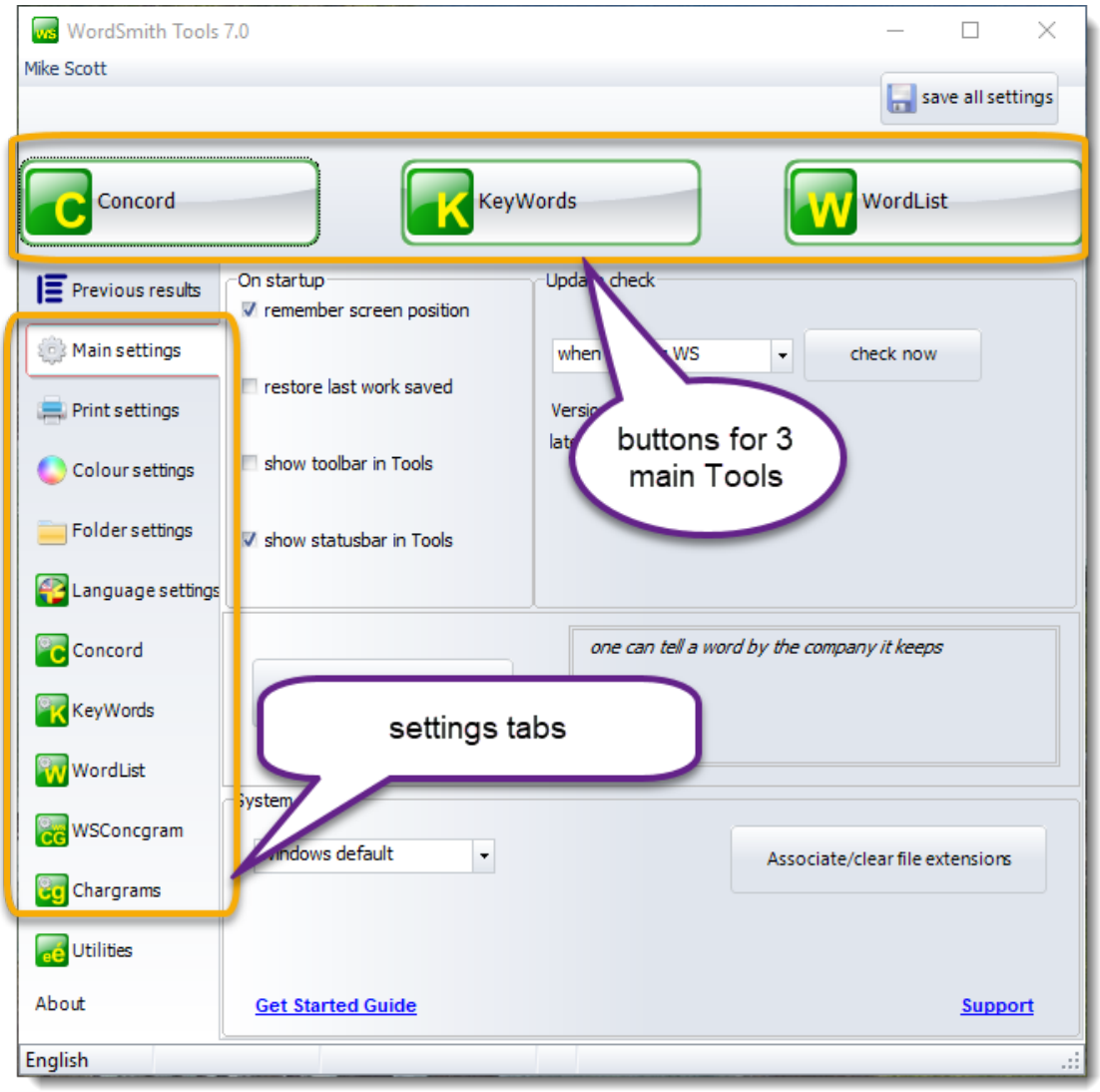
See also: version information 559, version updating 11.

Controller

Section V

5 Controller

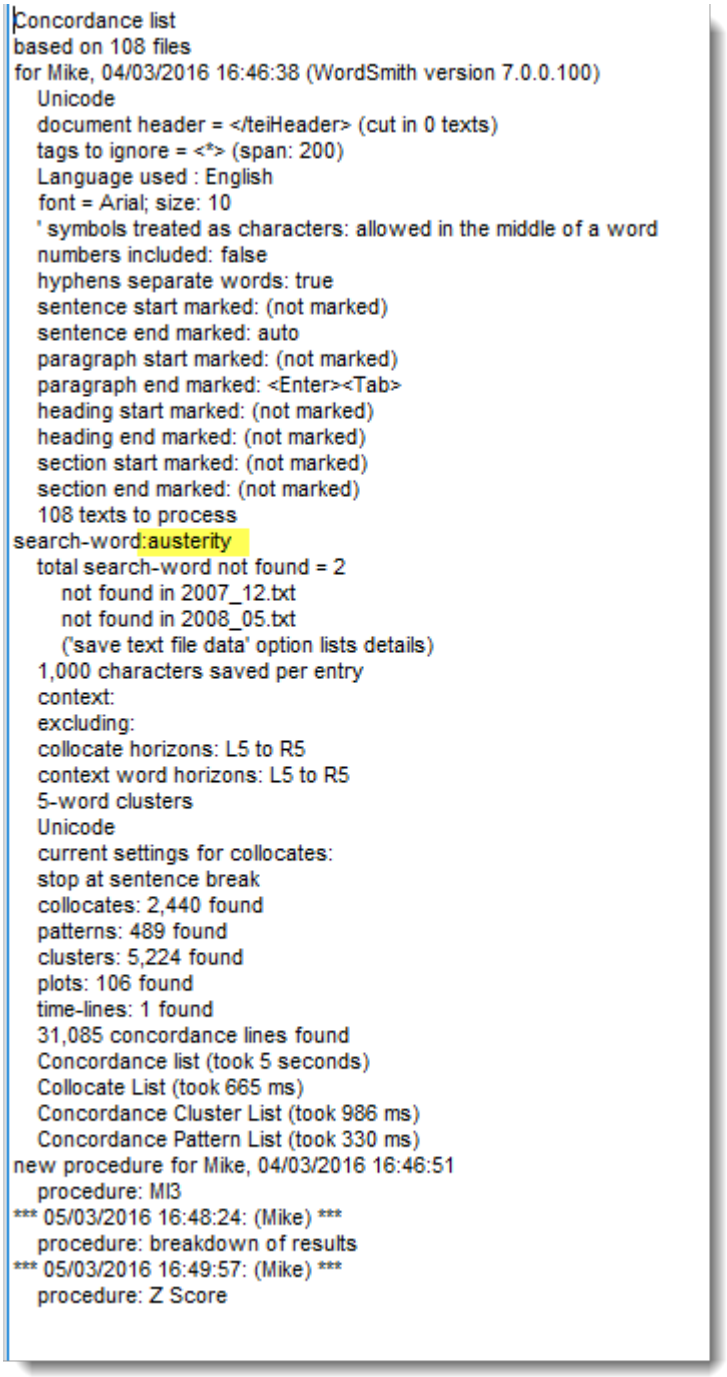
The main WordSmith Controller is a window which holds all the numerous settings and behind the scenes tells each Tool what to do. You can start up only one Controller -- though you can start up numerous Concord windows and WordList windows etc.



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

5.1 add notes

As WordSmith generates data, it will state the current relevant settings in the Notes tab and these are saved 242 with your data.



You may add to these notes, of course. For example, if you have done a concordance and sorted it carefully using your own user-defined categories 43, you will probably want to list these and save the information for later use.

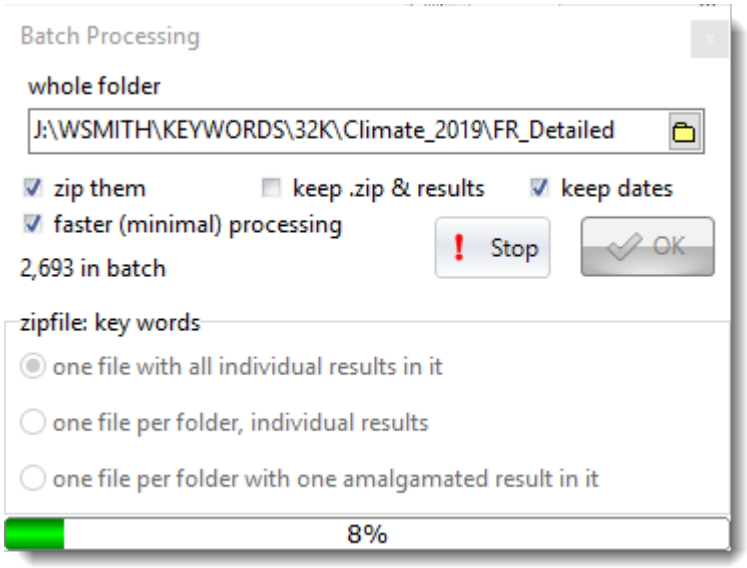
If you need access to these notes outside WordSmith Tools, select the text, then copy it to the clipboard 529 using Ctrl+C and paste into a word processor such as notepad.

5.2 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.



This screen shot shows a process which has already created over 2,000 key word lists. They are going to be stored in one large **.zip** file.

The maximum number of key word lists that can be created in a batch is 100,000.

folder where they end up

The name suggested is today's date[532]. 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\2020_07_19_12_00, they will end up like this:
c:\wsmith\wordlist\2020_07_19_12_00\0\fred1.lst
c:\wsmith\wordlist\2020_07_19_12_00\0\jim2.lst
..
c:\wsmith\wordlist\2020_07_19_12_00\0\mary512.lst
then
c:\wsmith\wordlist\2020_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** 561 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\2020_07_19_12_00\batch.zip

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

If you create more than 65,000 lists, they will get stored in multiple .zip files (**aaa.zip0, aaa.zip1 aaa.zip2** etc.)

One file / One file per folder?

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

But what if your text files are structured like this:

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

etc.

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

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

Batch Processing and Excel

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

faster (minimal) processing, keep .zip, keep dates, and results

The *faster (minimal) processing* 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 206, patterns 237 and dispersion plots 230 will be computed when each concordance is done. In KeyWords, likewise, there will be dispersion plots 296, link 292 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 537 don't get moved or deleted.

keep both .zip and results: if ticked, the **.KWS, .LST** or **.CNC** files created during the process won't get deleted after it ends, so you'll get both a **.zip** file and the expanded versions.

keep dates: If checked, preserves the file dates 149 of each list saved.

Example

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

You specify

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

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

After you press OK,

1,200 new word-lists are created, called **trunk.LST, tail.LST .. digestive system.LST**. They are all in numbered sub-folders of a folder called

C:\WSMITH\WORDLIST\ELEPHANT.

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

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

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

Getting at the results in WordSmith

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

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

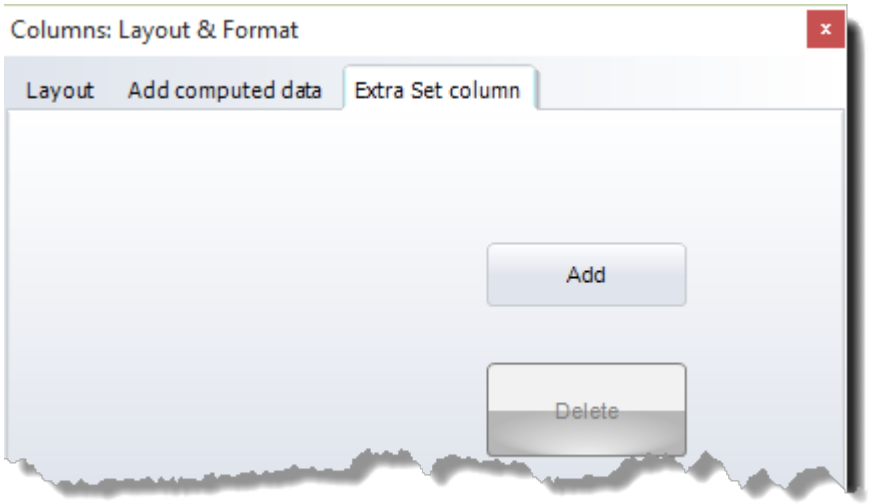
See also: batch scripts

5.3 category or Set

5.3.1 add a Set column

How to do it

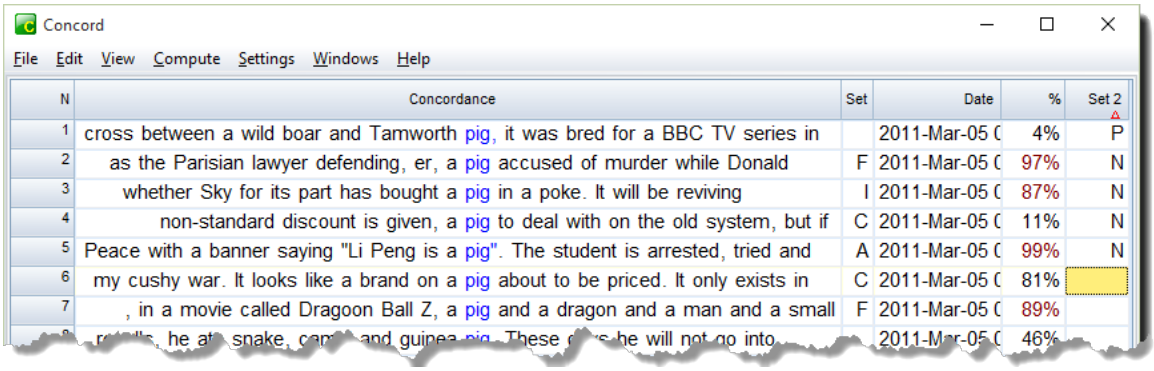
The Layout menu option (Ctrl + L) offers a tab headed *Extra Set column*.



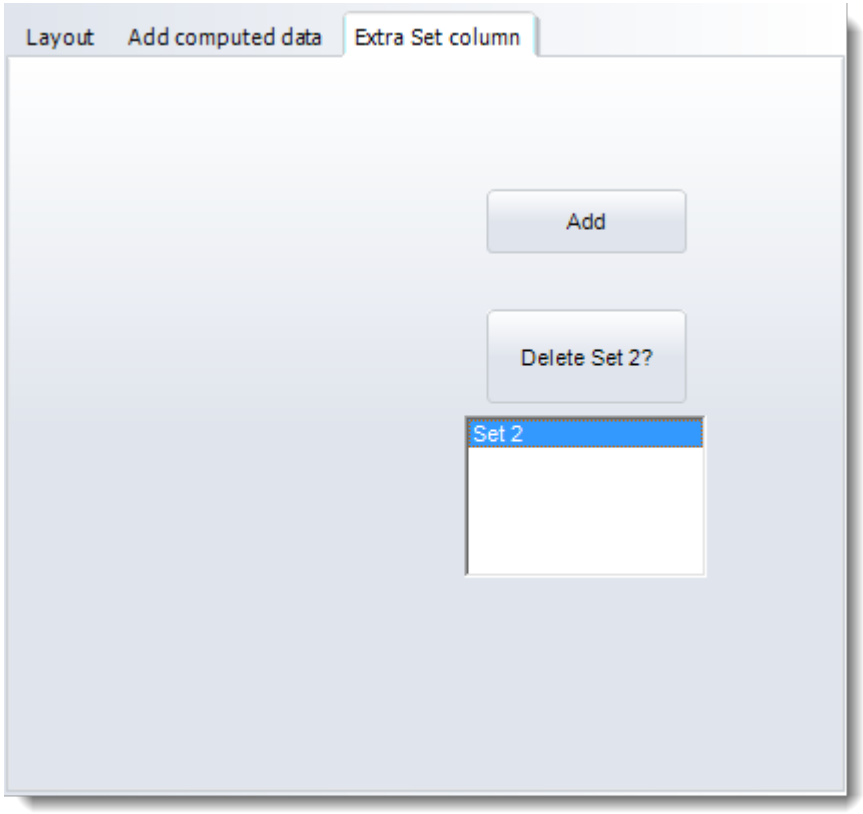
Here you can click to add a column (and delete any extra set columns created).

The point of it

is to aid in filtering and sorting your data. In this example, the word **pig** has been classified in the original set column according to whether it is Fictional, a Commodity, an Idiomatic usage, or Abuse. In the Set 2 column a further classification has been used to signal Positive or Negative contexts. It has been sorted on the Set 2 column.



Now it is possible to delete the additional Set column if desired.

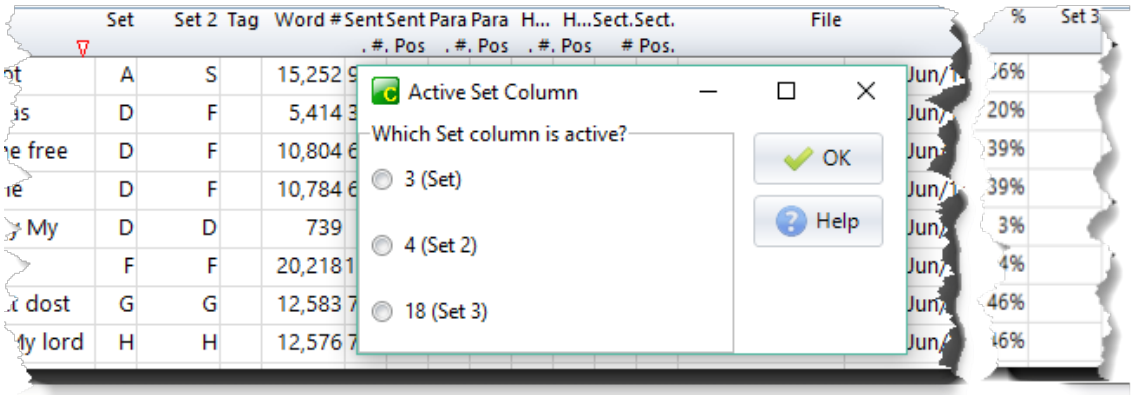


Using the new Set column

Click in the new column to select it, then enter your set column data 43 as desired.

Which Set column is active?

Use the menu (*Settings* | *Set column* | *Active column*) if you need to force your choice of which Set column gets the input as you type 43. Here there are three, in columns 3, 4 and (newly created, over at the right) 18. (You can move columns using Layout 60.)



Note: If you're typing in Set values for a row with multiple Set columns, the input will automatically go next to the next Set column on the same row before going down to the row below.

See also: compute new column 65, colour categories 35, type-in versus set 535 option, column layout 60.

5.3.2 colour categories

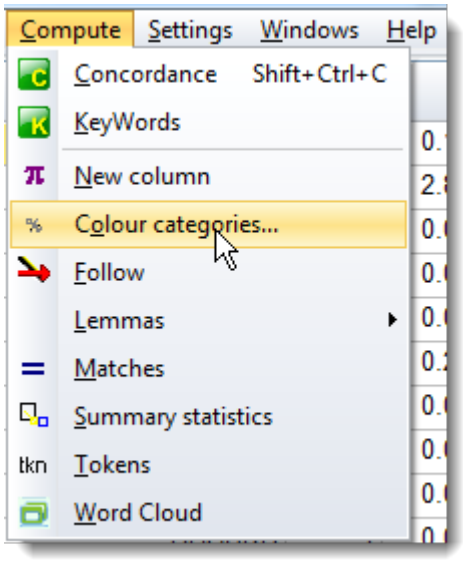
The point of it ...

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

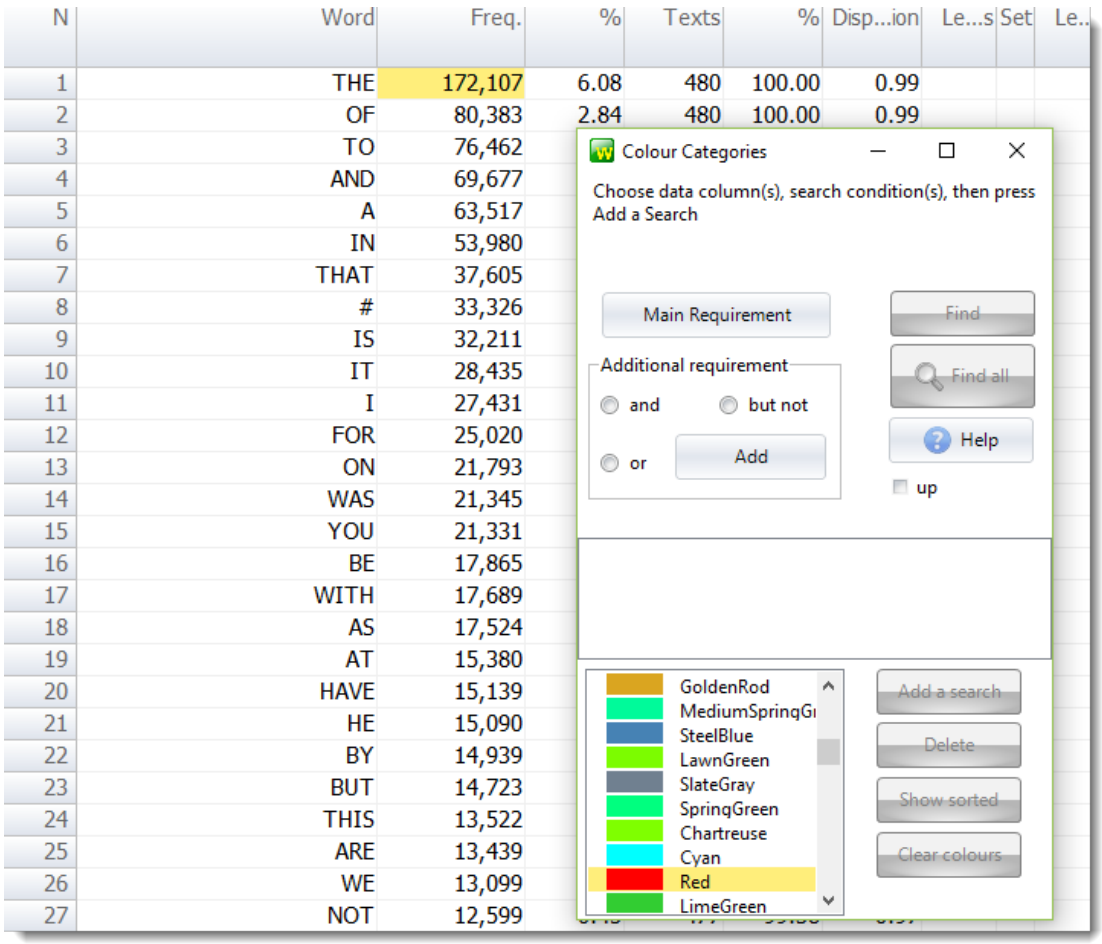
The idea is to let you re-sort 372 your existing data by your own criteria. (Since last millennium WordSmith lists have been sortable by standard criteria, and there has long been a Set 43 column for your own classification, but this feature makes it possible to have multiple and complex sorts.)

How to do it

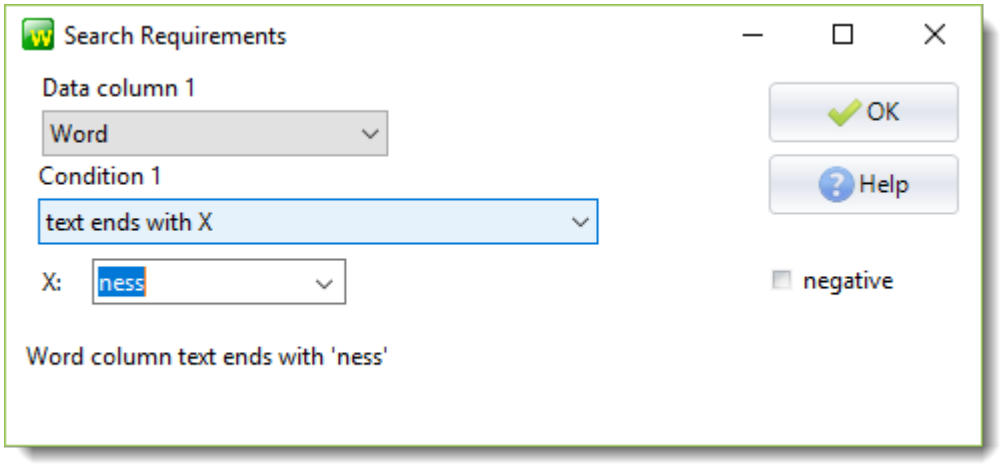
The menu option *Compute* | *Colour categories* will be found if the data have a Set column.



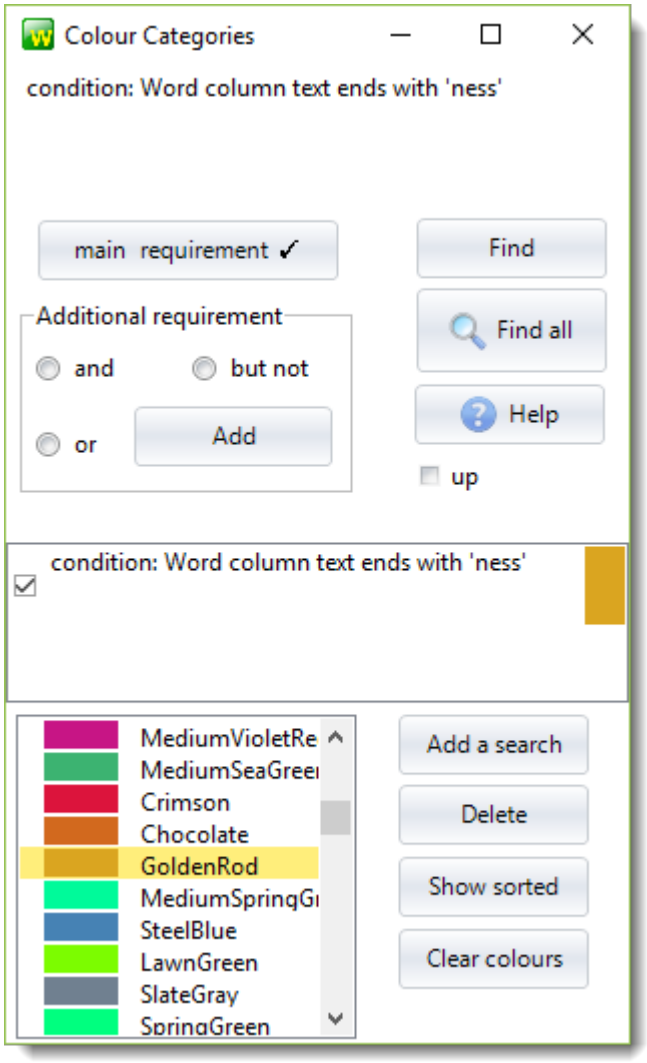
The menu option brings up a window where you specify your search criteria. Here is an example:



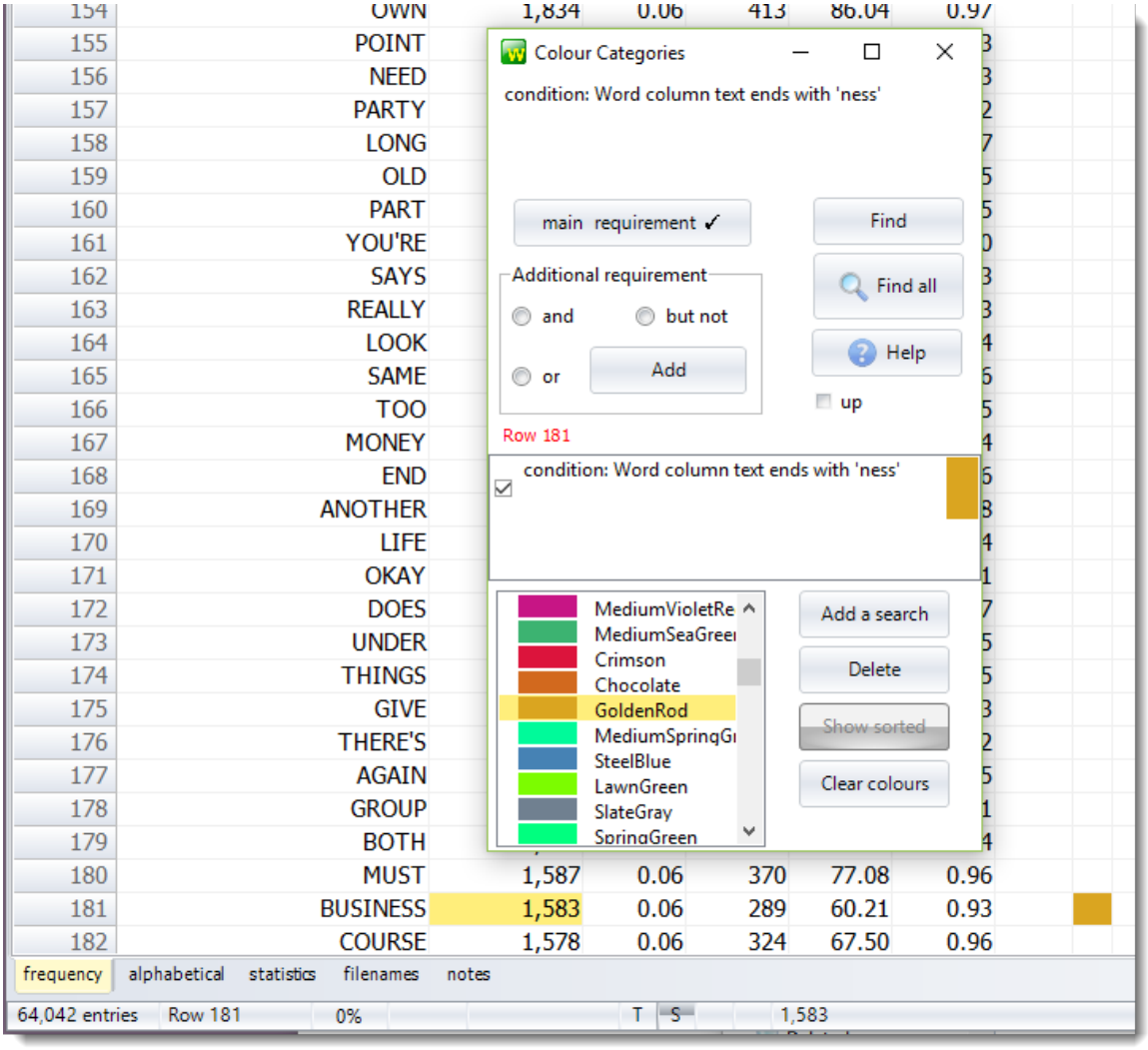
Complete the form by choosing a *Main Requirement*. Here, in *Data Column 1* the Word column was selected and the *text ends with* option. I typed in **ness**.



Now, having set a requirement, choose a colour to go with it and press *Add a Search*.



Finally, press *Find* or *Find All.* Here I pressed *Find*, and the first hit was **BUSINESS**, now coloured golden in the *Set* column.

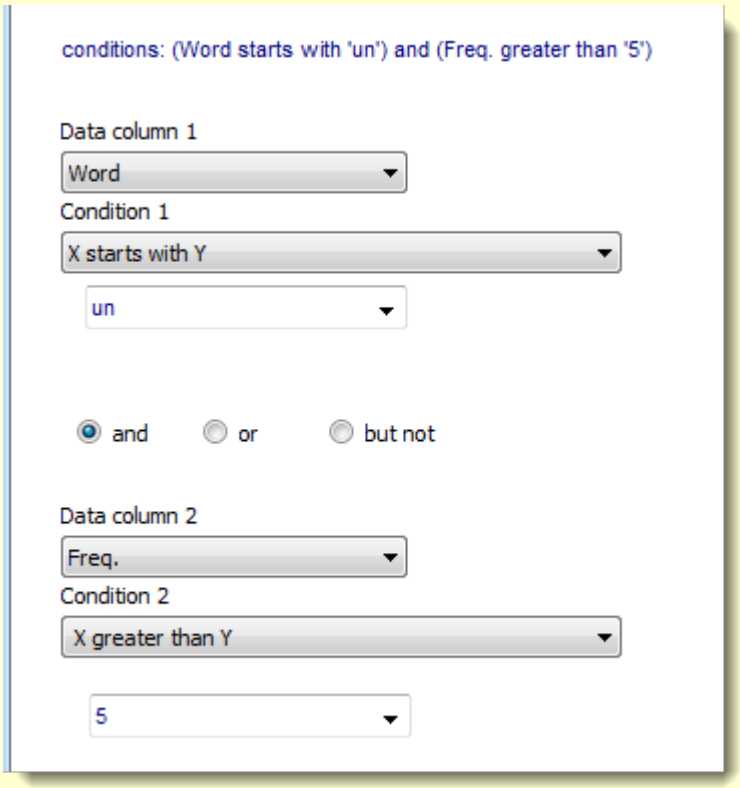


If you press *Find all* and sort the set column you can see all the items which met the requirement(s) set.

N	Word	Freq.	%	Texts	%	Di...on	Le...s	Set	Le...
1	WRETCHEDNESS	1		1	0.21				
2	WORTHLESSNESS	1		1	0.21				
3	WORTHINESS	2		2	0.42	0.35			
4	WORLDLINESS	6		6	1.25	0.58			
5	WONDROUSNESS	1		1	0.21				
6	WITNESS	104		56	11.67	0.78			
7	WISTFULNESS	1		1	0.21				
8	WINSOMENESS	1		1	0.21				
9	WILLINGNESS	39		32	6.67	0.84			
10	WILDERNESS	25		15	3.13	0.57			
11	WIDENESS	1		1	0.21				
12	WICKEDNESS	3		3	0.63	0.30			
13	WHOLESOMENESS	2		2	0.42	0.35			
14	WHITENESS	1		1	0.21				
15	WELSHNESS	1		1	0.21				
16	WEIRDNESS	2		2	0.42	0.35			
17	WEEDINESS	8		1	0.21				
18	WEATHEREDNESS	1		1	0.21				
19	WEARINESS	3		3	0.63	0.51			
20	WEAKNESS	46		41	8.54	0.92			
21	WATCHFULNESS	2		2	0.42	0.35			
22	WARINESS	1		1	0.21				
23	VIVIDNESS	2		2	0.42	0.35			

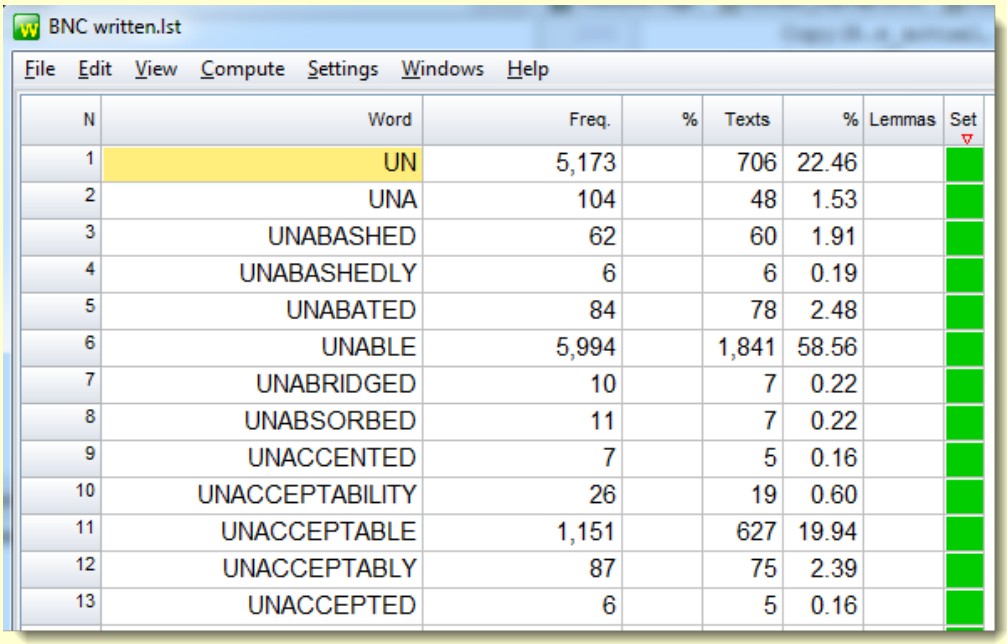
Advanced uses

A more complex example:



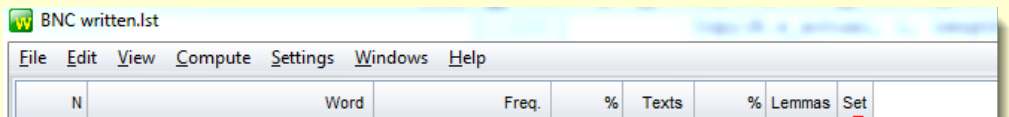
where the user wants to process the Word column of data, looking for a condition where the word starts with **UN** ***and*** occurs at least 5 times. For any word which meets this condition, the Set column will show the colour selected.

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



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

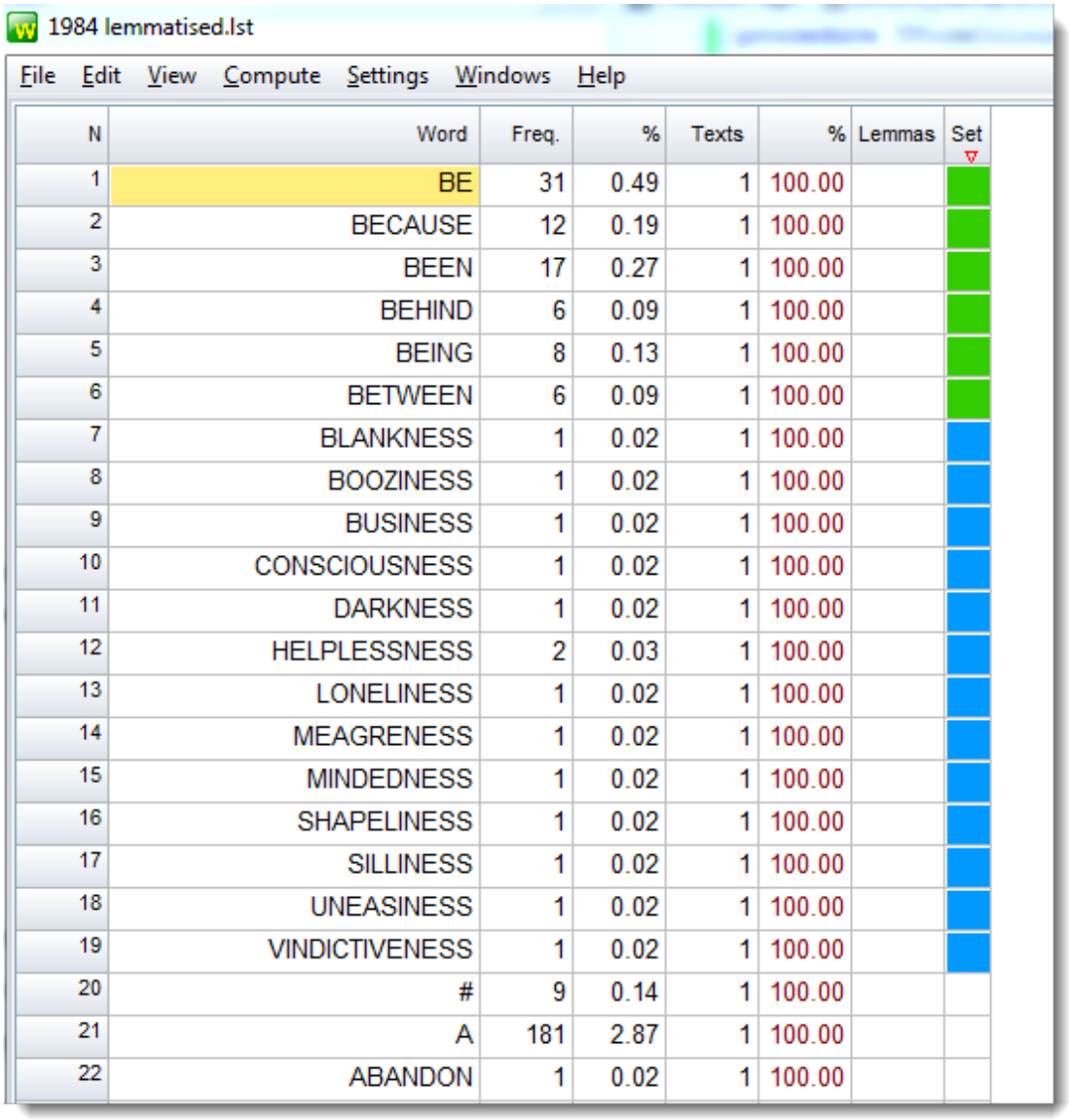
The top of the Colour Categories window shows the percentage results. In the example below, the user has decided to omit their first search and to carry out another on the same word-list which found 188 words ending in **NESS** which were present in more than 40 BNC texts.



N	Word	Freq.	%	Texts	%	Lemmas	Set

Where are they in the list?

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



1984 lemmatised.lst

File Edit View Compute Settings Windows Help

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

Once sorted, the data can be saved as before.

Removing the colours?

Use the Clear colours button.

See also : delete if 71, setting categories by typing 43, colour categories for concordances 45, search 95.

5.3.3 set column categories

The point of it…

You may want to classify entries in your own way, e.g. separating adjectival uses from nominal ones, or sorting according to different meanings.

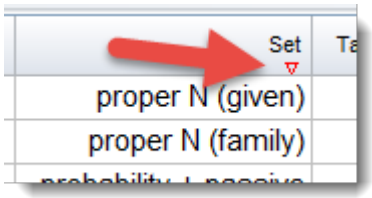
N	Concordance	Set
1	treasurer, David house manager, May, stage director, Norman play	proper N (given)
2	(Lens) □300,000. Out: David May (Manchester United) □1.25m,	proper N (family)
3	conditions have been met awards may be withheld. 14.2 Issue of	probability + passive
4	ruination of child prodigies? Girls may turn pro at 13 years 11	permission
5	and members of the Association, may I begin my report by thanking	permission
6	guaranteed by the EC budget may amount over 1990-2 to	P
7	four years behind schedule and may cost at least $;1 billion more	P
8	management and pessimism may be premature. .	P
9	in Letchmore Heath, however, may not prove that easy.	P
10	: be careful, because the Tutsis may come with their guns." "	P
11	it, "no point at which interpretation may reasonably stop" (1984: 20).	P
12	the heart of gangland. LeGeorge may be the best 12-and-under	P
13	culture in the last millennium may have had some 100 real	P
14	ownership laws no broadcaster may own more than 15 per cent of	P
15	, and began to deteriorate in May last year, May 1990. In the	M
16	expressed in the local elections of May 1990 and May 1991. The five	M

Here the user has used **P** where **may** has to do with probability and **M** if it's a month. In addition, some items have been labelled in more detail. You may wish also to modify your original texts 182 to include this annotation work you've done.

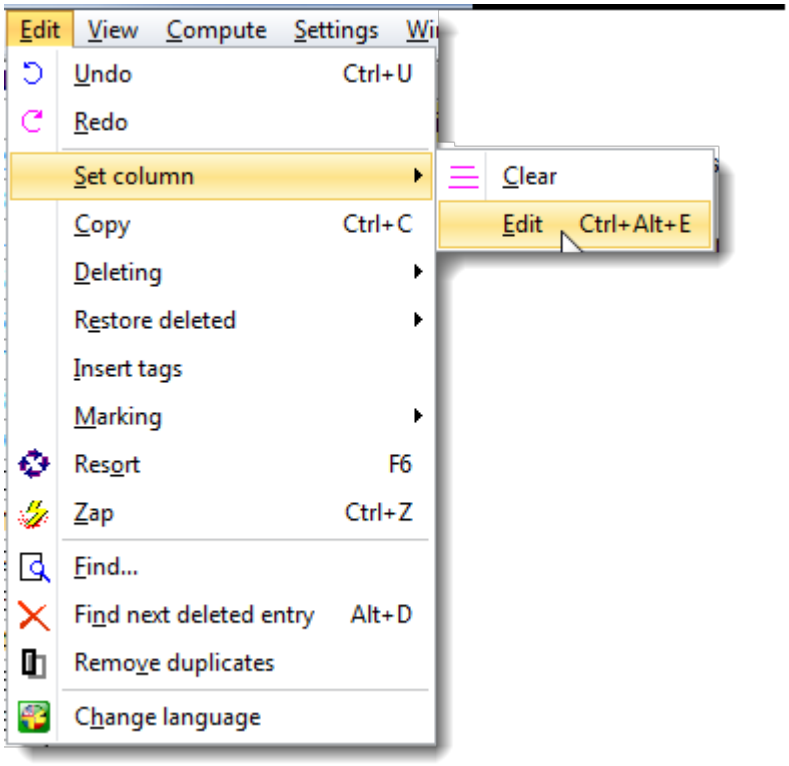
How to do it

Click once in the Set column. Then, if you simply press a letter or number key while the edit v. set v. type-in 535 mode is on Set you will get the concordance line marked with that letter or number in the Set column.

You can sort 238 the concordance lines using these categories, simply by clicking on the header Set, which will have a small triangle showing it's sorted.



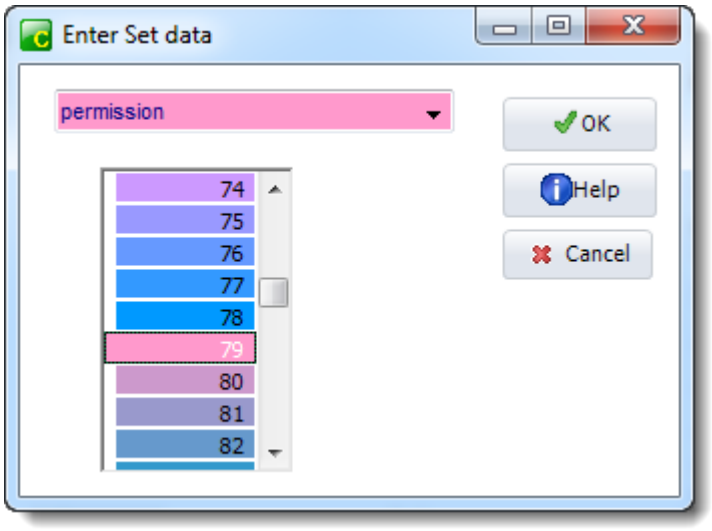
To enter the same value for various rows, first select the rows or mark 108 them, then choose *Set column | Edit*



then type in a suitable value.

More than 1 character? Colours

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



Here the word **permission** has been typed and the colour 79 has been dragged onto the box.

Clearing the Set column

To correct a mistake, press the zero key; that will remove any text and colour from the selected entry. (If you press the spacebar you will get blanking 199.)

Add another 33 Set column

See also : Copy search-word to Set column 252, Colour categories 35, modify your source texts 182, edit v. type-in 535 mode.

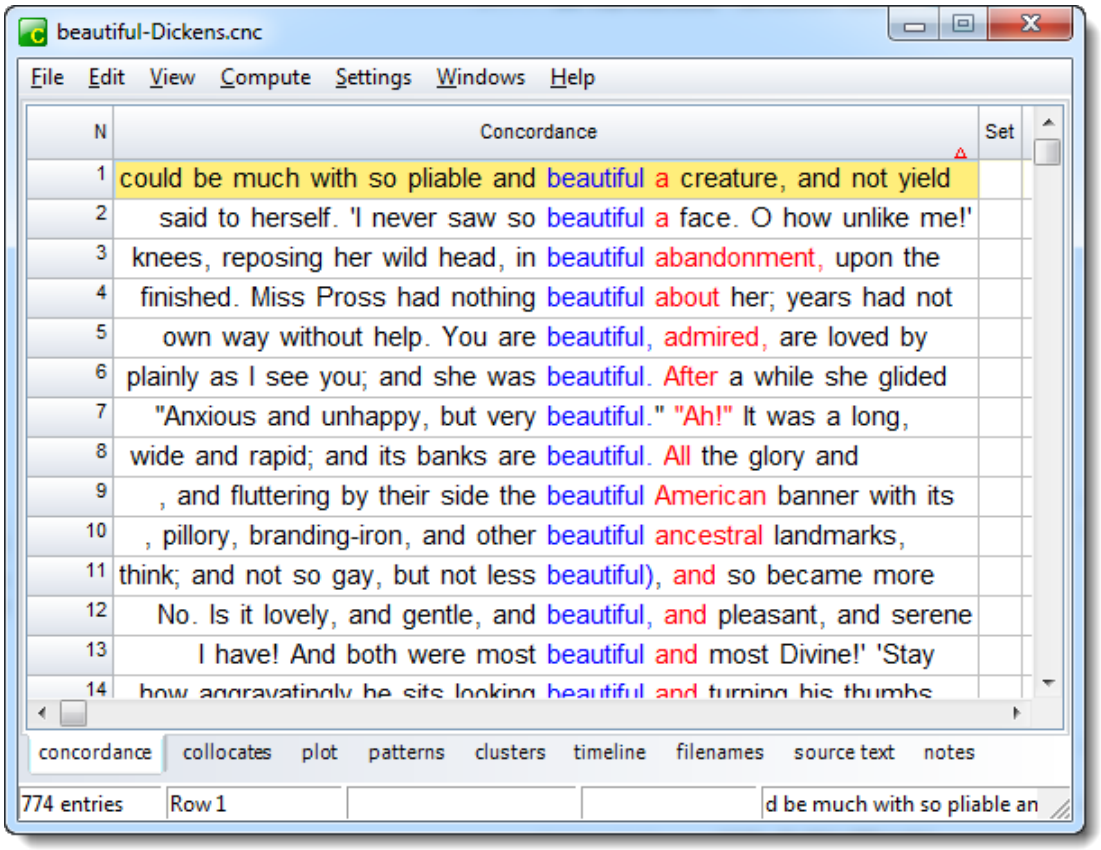
5.3.4 colour categories

The point of it...

The idea is to follow up a large concordance by breaking it down into specific sub-sections, so one can see how many of each sub-type are found in the whole list.

Example

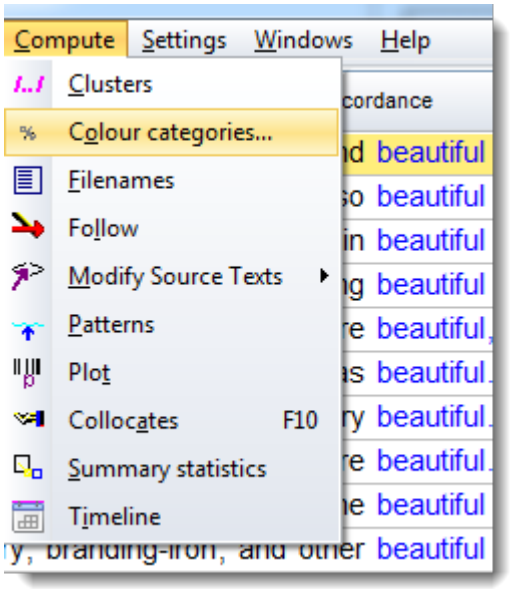
The screen-shot below came from a concordance of **beautiful** in Charles Dickens:



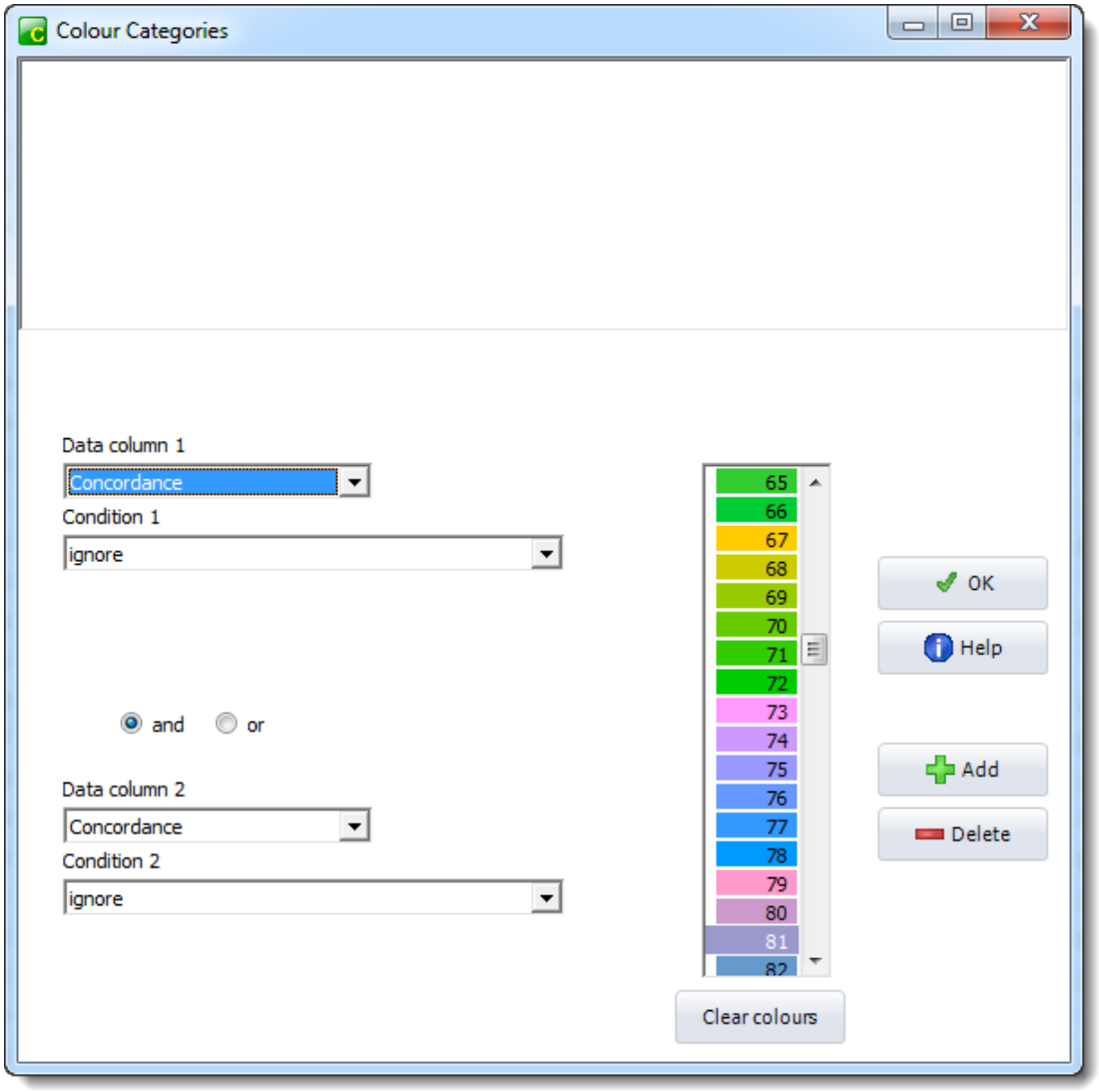
There are 774 lines. Looking through them, it became apparent that Dickens was fond of the collocations **`beautiful creature`** and **`beautiful face`**, but how many are of **`beautiful creature`** or similar (such as **`so.. beautiful a creature`** in line 1) and what proportion of the lines is that?

How to do it

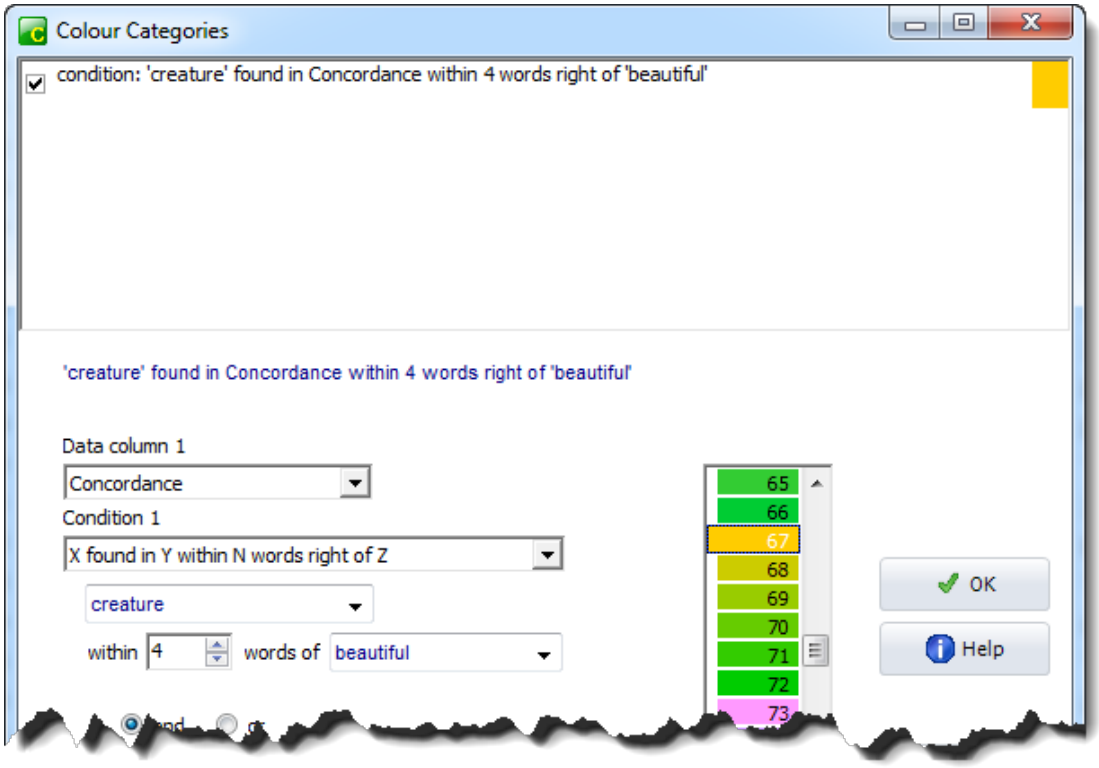
Choose *Compute* | *Colour Categories* in the menu.



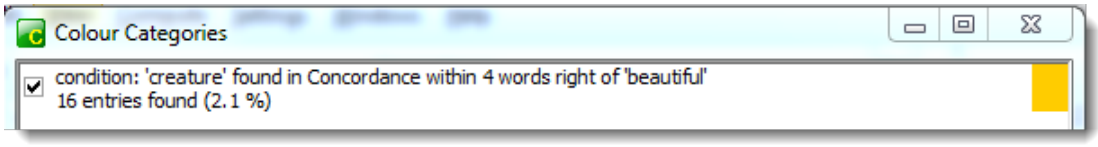
which opens up the colour categories box:



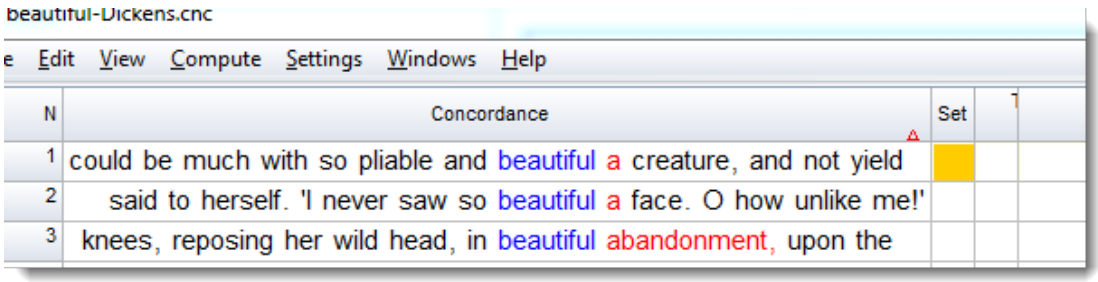
Here we have completed the spaces so as to get cases of beautiful ... with creature up to 4 words away to the right, and chosen



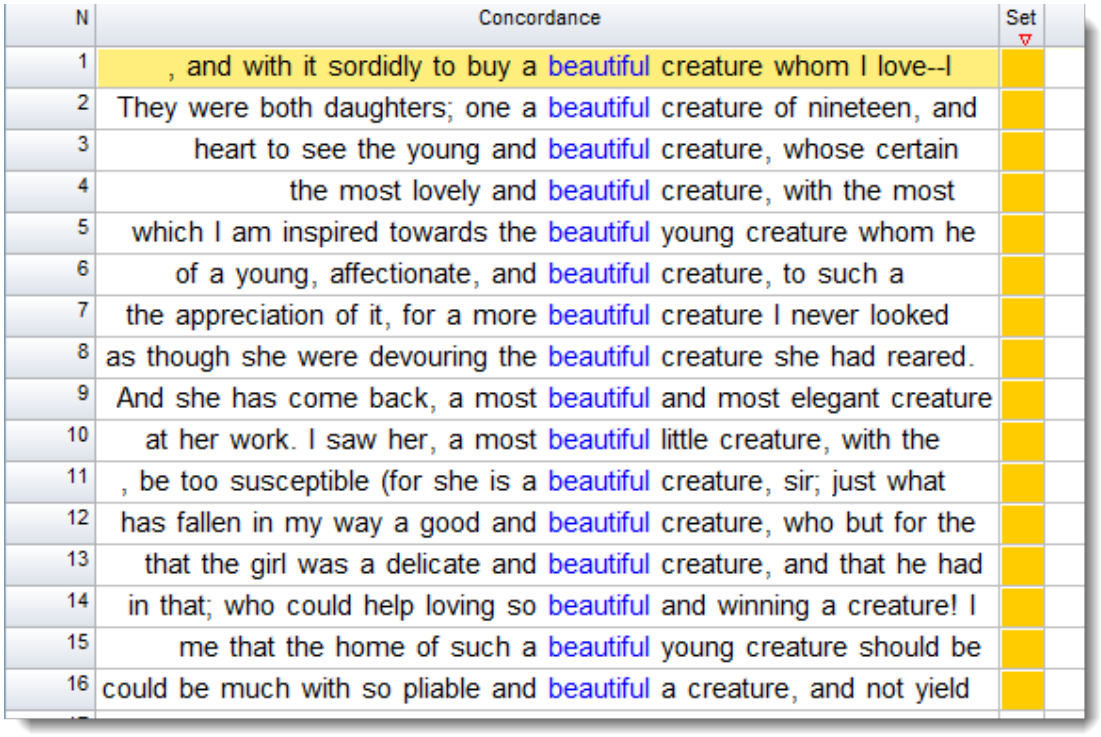
to colour yellow any which meet this condition. On pressing OK we find out there are 16, representing just over 2% of the lines.



and looking at the concordance the first line is now marked:



Where are the other 15? To find them, simply sort on the Set column.



N	Concordance	Set
1	, and with it sordidly to buy a beautiful creature whom I love--I	
2	They were both daughters; one a beautiful creature of nineteen, and	
3	heart to see the young and beautiful creature, whose certain	
4	the most lovely and beautiful creature, with the most	
5	which I am inspired towards the beautiful young creature whom he	
6	of a young, affectionate, and beautiful creature, to such a	
7	the appreciation of it, for a more beautiful creature I never looked	
8	as though she were devouring the beautiful creature she had reared.	
9	And she has come back, a most beautiful and most elegant creature	
10	at her work. I saw her, a most beautiful little creature, with the	
11	, be too susceptible (for she is a beautiful creature, sir; just what	
12	has fallen in my way a good and beautiful creature, who but for the	
13	that the girl was a delicate and beautiful creature, and that he had	
14	in that; who could help loving so beautiful and winning a creature! I	
15	me that the home of such a beautiful young creature should be	
16	could be much with so pliable and beautiful a creature, and not yield	

This function applies to word lists and other data too, and is explained in more detail the main colour categories 35 section. The set column itself can contain characters or words as well as colours, as explained in the set column 43 section.

5.4 choosing texts

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

You need text in a suitable format 49. There is also an explanation of the file-choosing window 52 and a way of saving your selection as a favourite text 59 selection.

5.4.1 text formats

In WordSmith you need plain text files 527, such as you get if you save a Word .doc 552 as Plain Text (**.txt**). The text format should be ASCII or ANSI or Unicode (UTF16). HTML and XML are OK too.

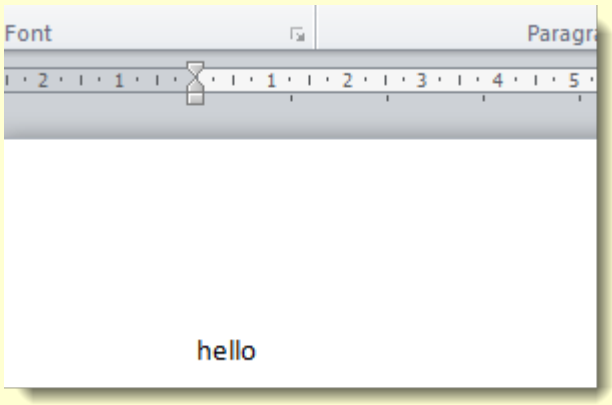
Any Word **.doc** or **.docx** files will look crossed out and should not be used: convert them to .txt first 552.

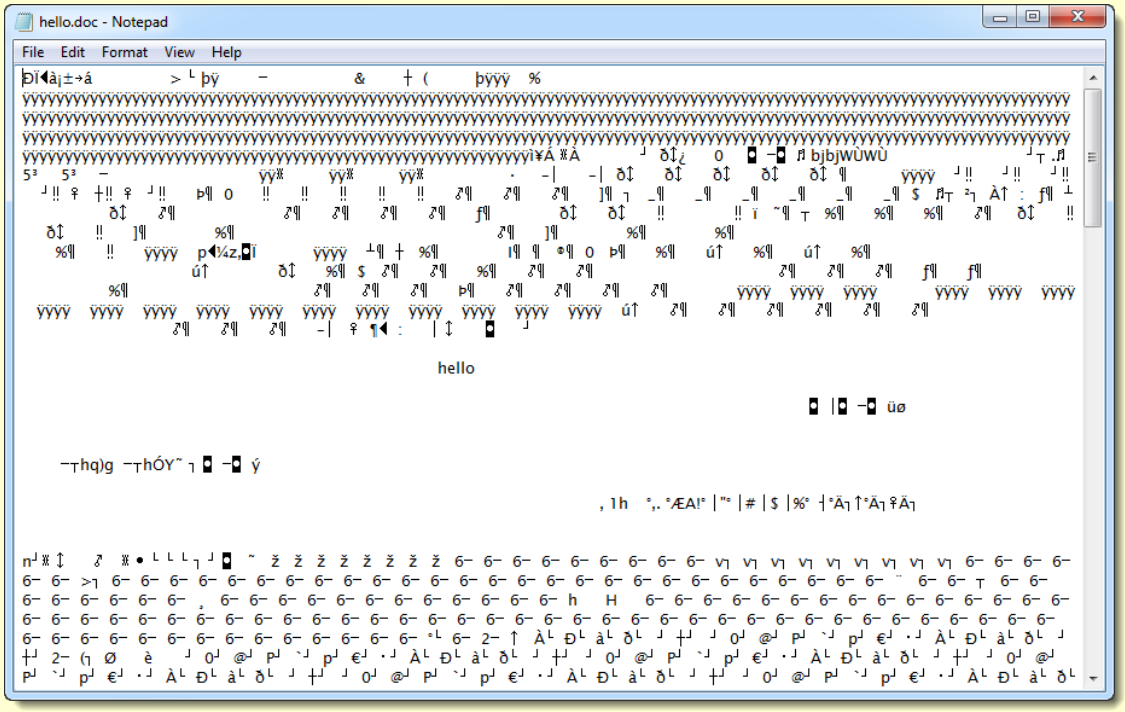
Text files within .zip files 58 can be used; they will be coloured red in the *Files available* display but WordSmith can read them and get the texts you select within them.

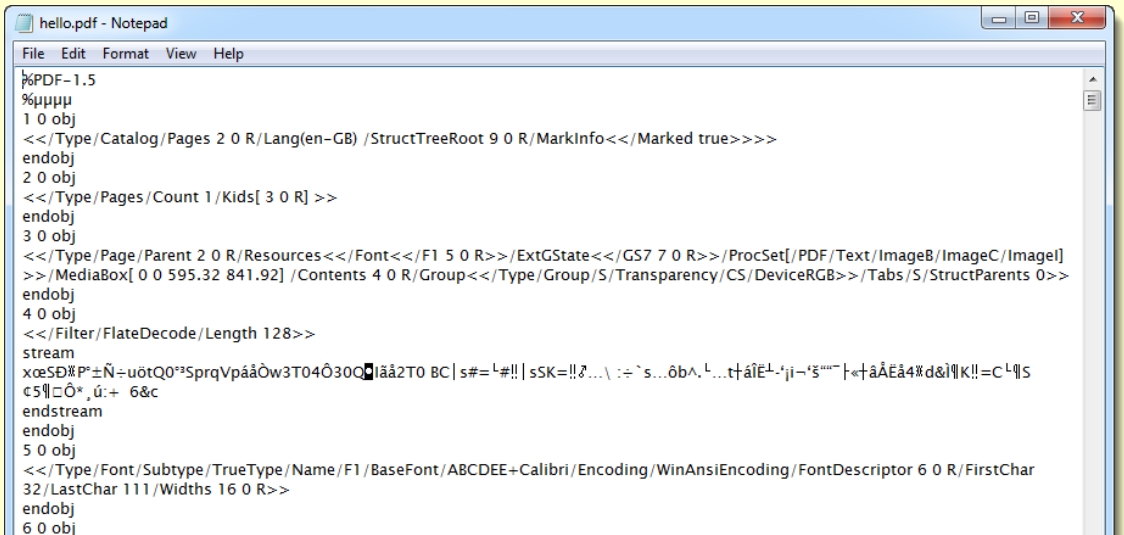
Why not .PDF or .DOC files?

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

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







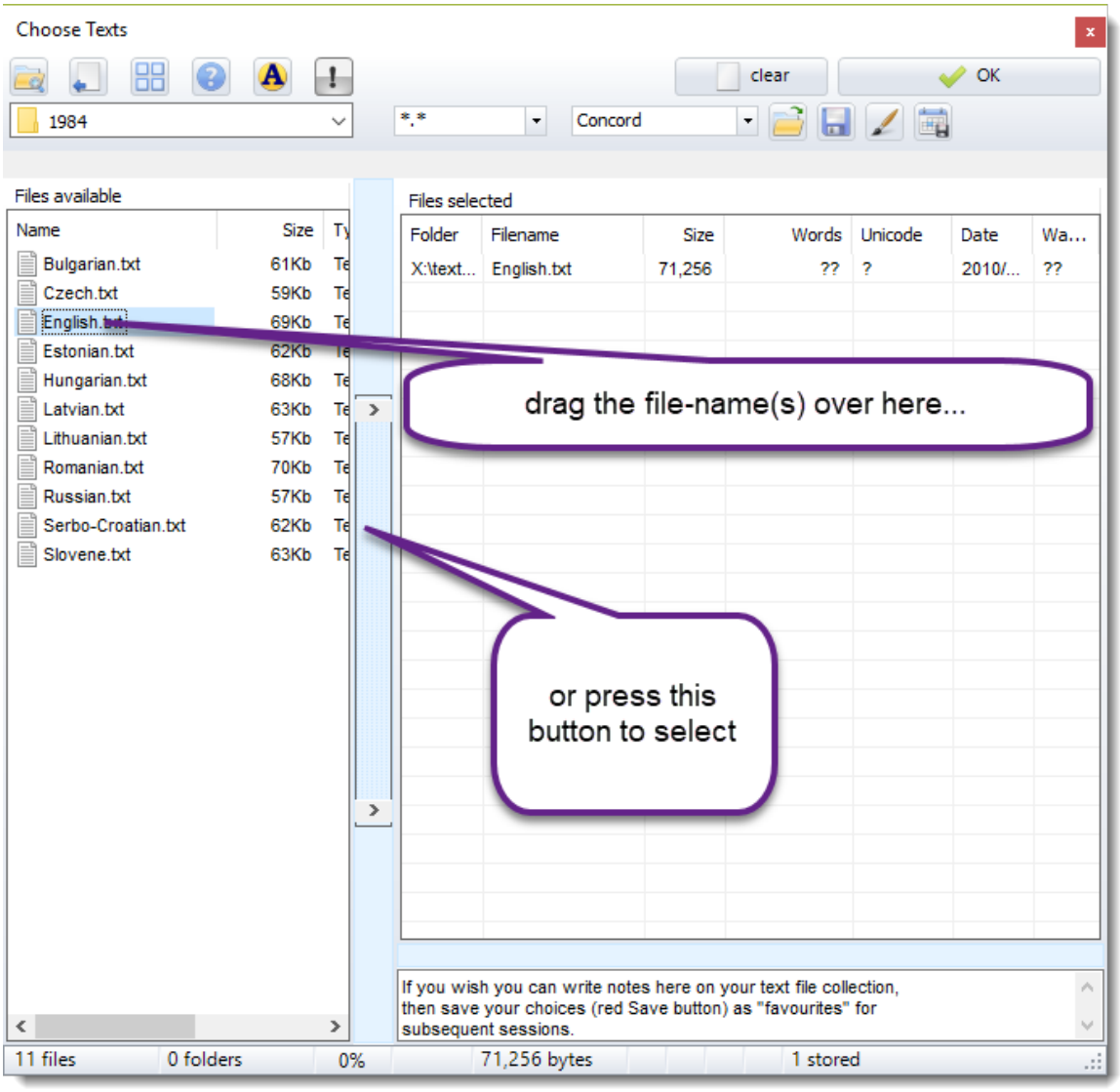
Check the format is OK

In the file-choose window you can test the format of the texts you've chosen with the *Test File Format* 54 ( ) button.

5.4.2 the file-choose window

How to get here

This function is accessed from the File menu in the Controller 5 and the Settings menu or New menu item ( ) in the various Tools.



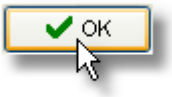
The two main areas at left and right are

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

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

The list on the right shows full file details (name, date, size, number of words (above shown with ?? as WordSmith doesn't yet know, though it will after you have concordanced or made a word list) and whether the text is in Unicode (? for the same reason). To the right of Unicode is a column stating whether each text file meets your requirements 171.

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 87.



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

Clear

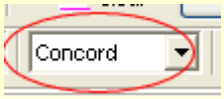
As its name suggests, this allows you to change your mind and start afresh. If any selected filenames are highlighted, only these will be cleared.

More details

File Types

The default file specification is *.* (i.e. any file) but this can be altered in the box or set permanently in wordsmith7.ini 110.

Tool

In the screenshot above you can see 

-- we are choosing texts for Concord. There are alternatives available (WordList, KeyWords etc.).

Select All

Selects all the files in the current folder.

Drives and Folders

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

Sub-Folders

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

Sorting

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

Test text format (Unicode test)

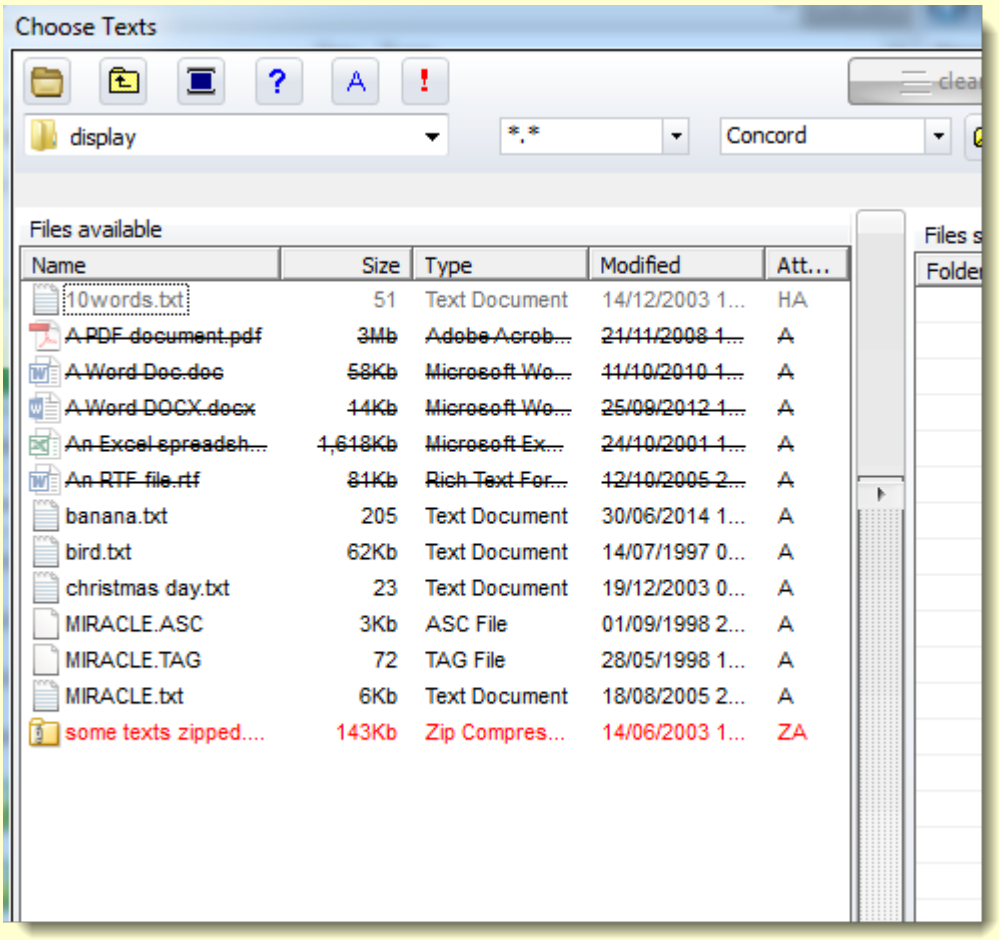
This button checks the format of any files selected. In the screenshot above, no tests have been done so the display shows ? for each file. If the text file is in Unicode, the display shows *U*, if Unicode big-endian it'll show *UB*, if plain ASCII or Ansi text it will show *A*, if it's a Word .doc file,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

Favourites 59

Two buttons on the right (and) allow you to save or get a previous file selection 59, saving you the trouble of making and remembering a complex set of choices.

Type of text files

In WordSmith you need plain text files[527], such as you get if you save a **Word .doc**[552] as Plain Text (**.txt**). Any Word **.doc** files or **.pdf**s will look crossed out and should not be used: convert them to .txt first[552].



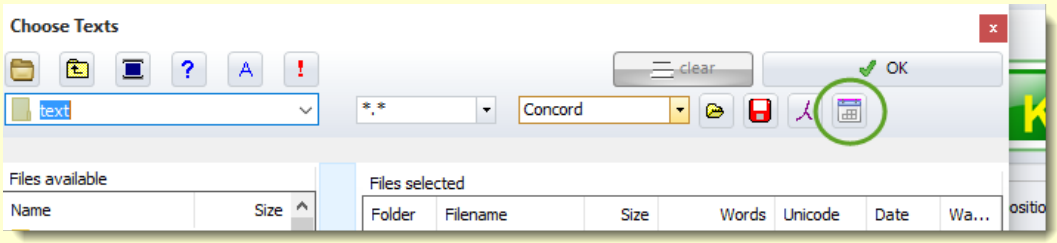
(*10words.TXT* is greyed out because it has the hidden attribute)

Setting text file dates

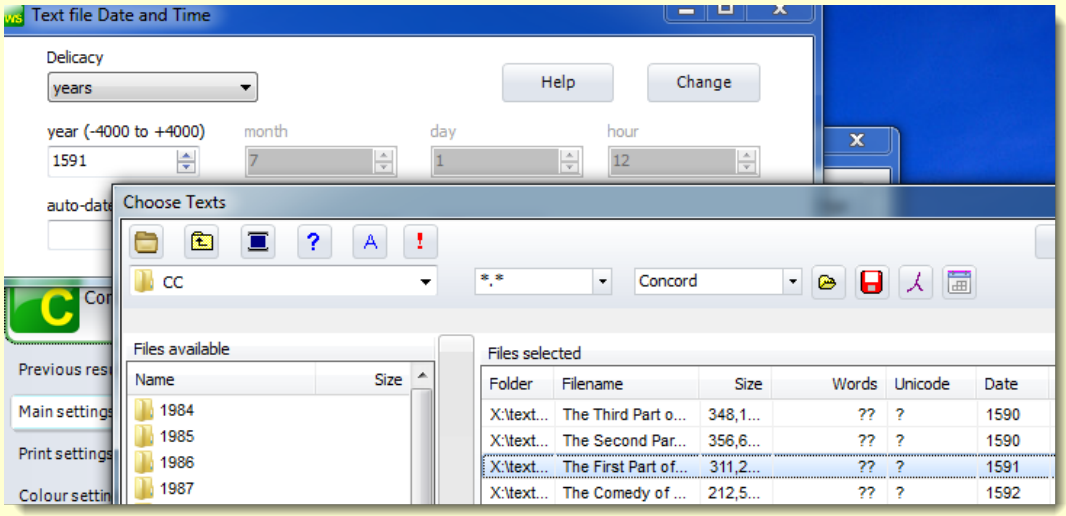
You can edit the textual date to be attached to any text file within any date range from 4000BC upwards. (On first reading from disk the date will be set to the date that text file was last edited.)

How to do it

Press the button circled in this screen-shot:

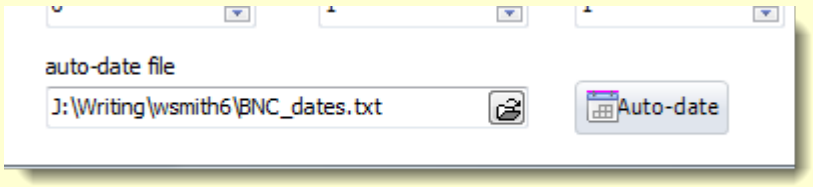


A window opens up letting you set text file dates and times. Here below you will see Shakespeare plays with their dates being edited.



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

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



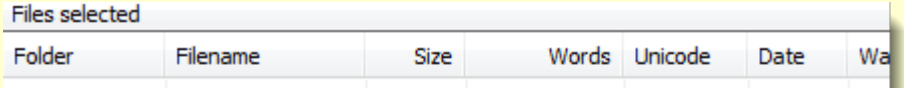
The format of the list is

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

Examples:

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

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



Advanced

Opens a toolbar showing some further buttons:

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

Random

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

View in Notepad

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

Check

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

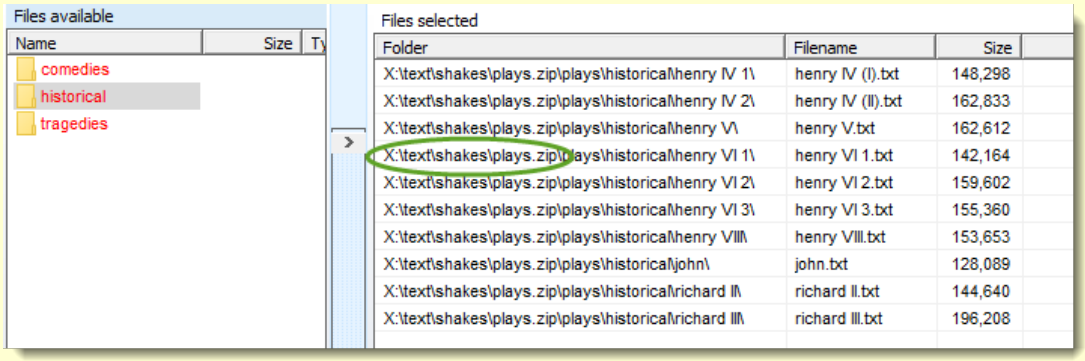
Save List .txt

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

Zip files

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

Whether checked or not, if you double-click on a zip file 561 you can enter that as if it were a folder and see the contents. Zip files will be coloured red. In this screen, the historical plays of Shakespeare within a zip file (**`plays.zip`**) have been selected.



See also : Step-by-step online example, Viewing source texts 134, Finding source texts 537.

5.4.3 favourite texts

save favourites

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

Saves a list of text files whose status is either unknown or known to meet your requirements when selecting files by their contents 171, ignoring any which do not.

get favourites

Used to read a previously-saved selection from disk.

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

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

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

See also: Choosing Texts 52, file dates 56

5.5 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 **\wsmith7** folder where the WordSmith programs are or in the "instructions folder" explained under site licence defaults 25. If one is found, it will be shown automatically when WordSmith starts up. To stop it being shown, just rename it! You edit the file using any Rich Text Format word processor, such as MS Word™, saving as an **.rtf** file.

See also: Site licence defaults 25

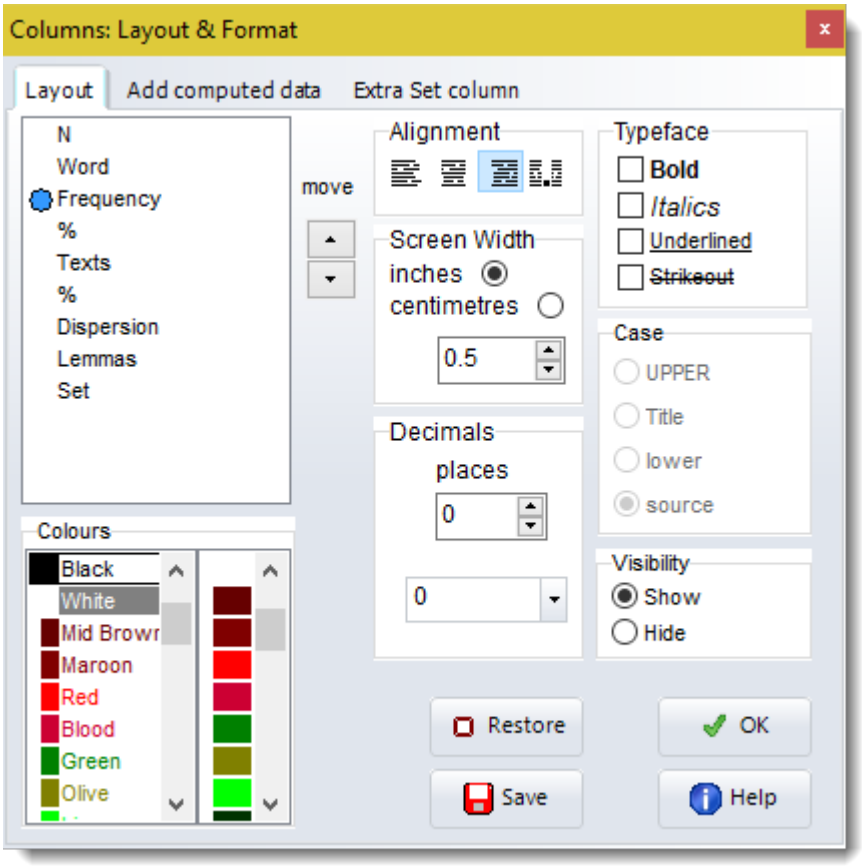
5.6 columns of data

5.6.1 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 65. And you can add an extra set column 33 if you like.



You can edit the headings 74 by double-clicking and typing in your own preferred heading. Here I've changed *Freq.* to *Frequency*.

Move

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

Alignment

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

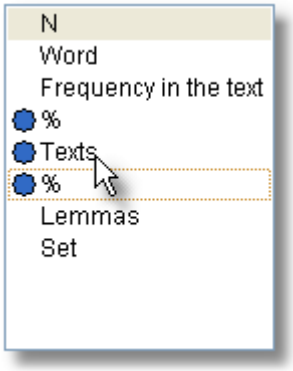
Typeface

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

Screen Width

in your preferred units (cm. or inches).

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

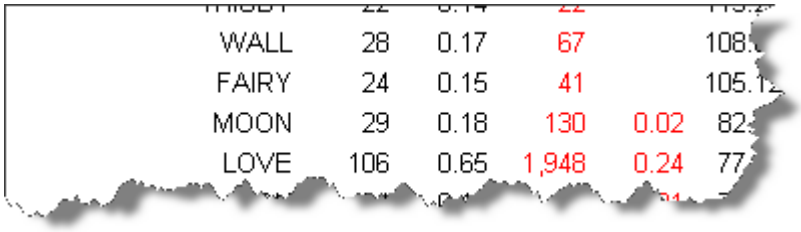


Case

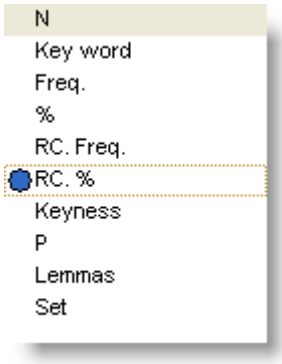
lower case, UPPER CASE, Title Case or source: as it came originally in the text file. The default for most data is upper case.

Decimals

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

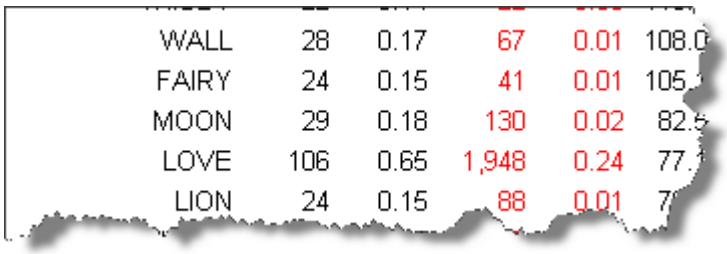


select the column(s) you want to affect,



and set the decimals eg. like this

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



Visibility

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

Colours

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

Restore

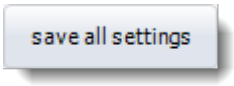
Restores settings to the state they were in before. Offers a chance to delete any custom saved layout for the current type of data (see Save).

Save

The point of this Save option is to set all future lists of the same type as the one you're working on to a preferred layout. Suppose you have a concordance open. If you change the layout as you like and save 91 the concordance in the usual way it will remember your settings anyway. But the next time you make a concordance, you'll get the WordSmith default layout. If you choose this Save, the next time you make a concordance, it will look like the current one.

And a custom saved layout will be found in your **Documents\wsmith7** folder, eg. **Concordance list customised.dat. (**The only way of removing such settings would be to rename or delete that file.)

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



which can be saved permanently with .

Freeze columns

If you have a lot of detailed consistency files and wish to freeze the word column so as to see the words for every column of numbers, choose View | Freeze columns... This allows you to freeze the number of fixed columns, which lets you scroll right to see data more clearly:

Here there are 480 text files and you can easily see the total & number of texts for text file **A00, A01, A0A** and so on,

File Edit View Compute Settings Windows Help

N	Word	Total	Texts	No. of Lemmas	Set	A00	A01	A0A
1	THE	172,107	480	0		315	363	475
2	OF	80,383	480	0		198	233	180
3	TO	76,462	480	0		208	289	167
4	AND	69,677	480	0		201	174	199
5	A	63,517	480	0		134	206	176
6	IN	53,980	480	0		132	118	129
7	THAT	37,605	479	0		30	63	35
8	#	33,326	430	0		166	163	99
9	IS	32,211	480	0		69	113	134

but when you have scrolled right to see data on texts **A7V, A7Y** you can't see the words or their totals.

N	A7V	A7Y	A80	A8
1	642	663	686	40
2	279	317	216	22
3	242	316	252	17
4	182	293	206	21
5	178	261	208	17
6	228	192	201	14
7				

The solution is to freeze some columns. Here I froze 4 columns:

N	Word	Total	Texts	A7V	A7Y	A80
1	THE	172,107	480	642	663	686
2	OF	80,383	480	279	317	216
3	TO	76,462	480	242	316	252
4	AND	69,677	480	182	293	206
5	A	63,517	480	178	261	208
6	IN	53,980	480	228	192	201
7	THAT	37,605	479	108	127	71

The frozen columns (which cannot be re-sorted without setting freeze back to 1) displays with smaller font. Noe you can easily look at data for text files **A7V, A7Y** etc..

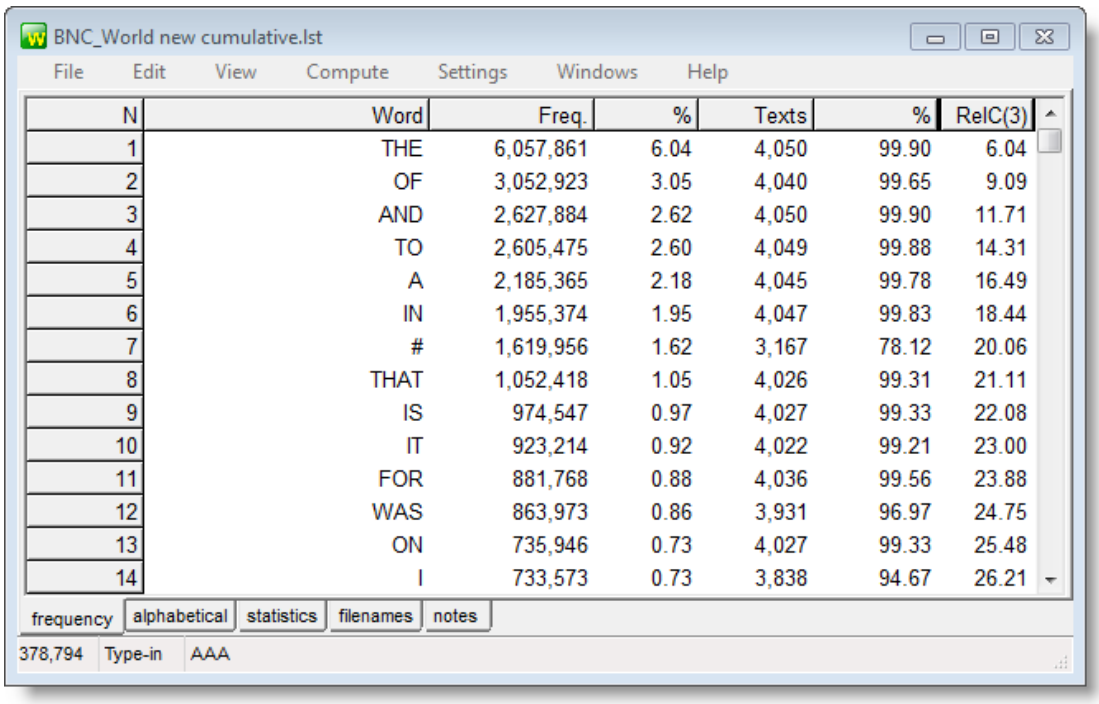
Similarly a statistics list allows the text file-names column to be frozen:

See also: percentage_columns 131, setting & saving defaults 110, setting colour 114 choices in WordSmith Tools Controller 5.

5.6.2 compute new column of data

The point of it…

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

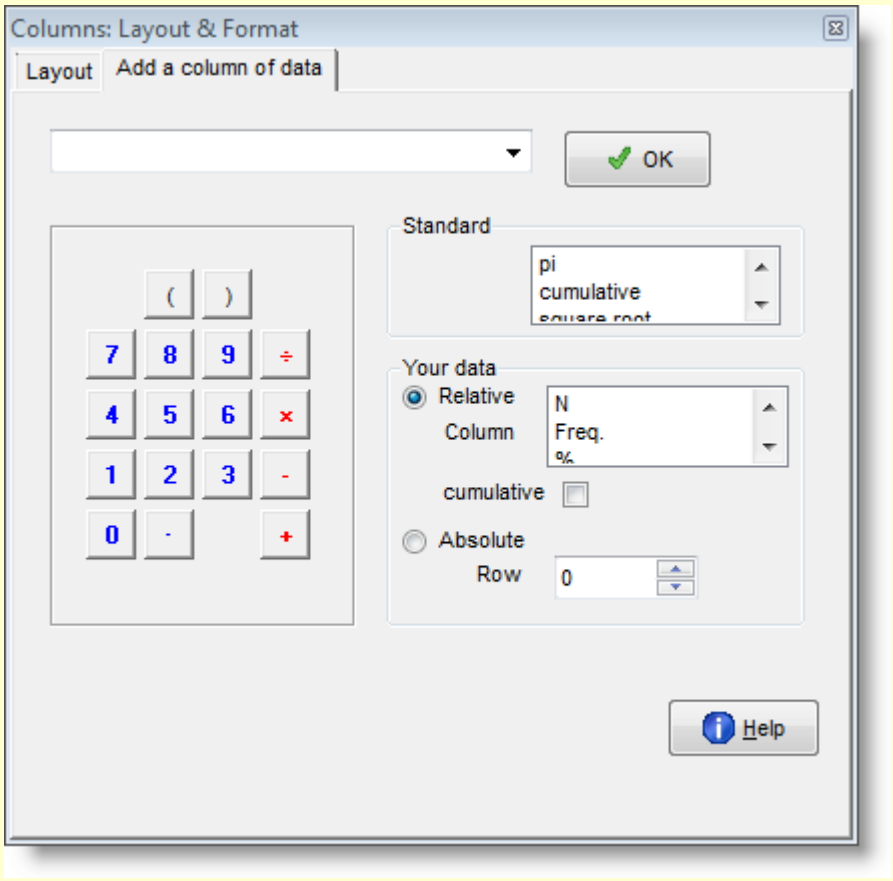


N	Word	Freq.	%	Texts	%	RelC(3)
1	THE	6,057,861	6.04	4,050	99.90	6.04
2	OF	3,052,923	3.05	4,040	99.65	9.09
3	AND	2,627,884	2.62	4,050	99.90	11.71
4	TO	2,605,475	2.60	4,049	99.88	14.31
5	A	2,185,365	2.18	4,045	99.78	16.49
6	IN	1,955,374	1.95	4,047	99.83	18.44
7	#	1,619,956	1.62	3,167	78.12	20.06
8	THAT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

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

How to do it

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



Absolute and Relative

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

Examples

you type	Result -- for each row in your data, the new column wil contain:
Rel(2) ÷ 5	the data from column 2 of the same row, divided by 5
RelC(2)	the data from column 2 of the same row, added to a running total
Rel(3) + (Rel(2) ÷ 5)	the data from column 2 of the same row, divided by 5, added the data from column 3 of the same row
Abs(2;1) ÷ 5	the data from column 2 of row 1, divided by 5. (This example just to illustrate; it would be silly as it would give the exact same result in every row.)
Rel(2) ÷ Abs(2;1) × 100	the data from column 2 of the same row divided by column 2 row 1 and multiplied by 100. This would give column 3 as a

percentage of the top result in column 2. For the first row it'd give 100%, but as the frequencies declined so would their percentage of the most frequent item.

You can format (or even delete) any variables computed in this way: see layout 60.

See also: count data frequencies 141, column totals 67, colour categories 35, add a Set column 33

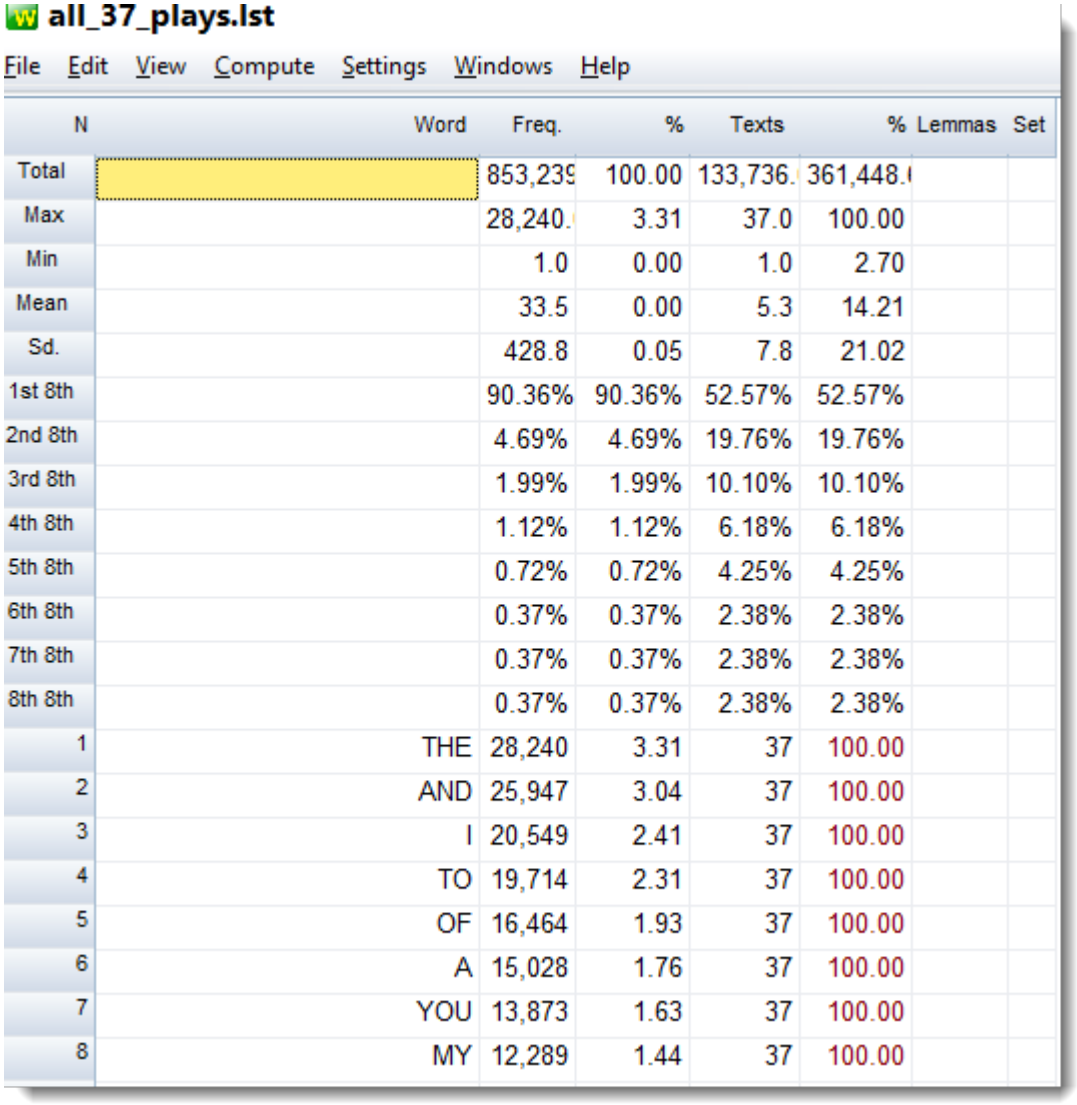
5.6.3 column totals

The point of it...

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

How to do it

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



all_37_plays.lst

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%	Lemmas	Set
Total		853,239	100.00	133,736.	361,448.		
Max		28,240.	3.31	37.0	100.00		
Min		1.0	0.00	1.0	2.70		
Mean		33.5	0.00	5.3	14.21		
Sd.		428.8	0.05	7.8	21.02		
1st 8th		90.36%	90.36%	52.57%	52.57%		
2nd 8th		4.69%	4.69%	19.76%	19.76%		
3rd 8th		1.99%	1.99%	10.10%	10.10%		
4th 8th		1.12%	1.12%	6.18%	6.18%		
5th 8th		0.72%	0.72%	4.25%	4.25%		
6th 8th		0.37%	0.37%	2.38%	2.38%		
7th 8th		0.37%	0.37%	2.38%	2.38%		
8th 8th		0.37%	0.37%	2.38%	2.38%		
1	THE	28,240	3.31	37	100.00		
2	AND	25,947	3.04	37	100.00		
3	I	20,549	2.41	37	100.00		
4	TO	19,714	2.31	37	100.00		
5	OF	16,464	1.93	37	100.00		
6	A	15,028	1.76	37	100.00		
7	YOU	13,873	1.63	37	100.00		
8	MY	12,289	1.44	37	100.00		

This view shows the frequency list of Shakespeare plays with column totals. In the % column we see that the maximum was 3.31%. In other words, the top frequency word (**THE**) takes up 3.31% of the running words in these plays. The mean *Freq*. is 33.5 and the mean number of plays is 5.3. In other words the average word in this set of data crops up in about 5 of the Shakespeare plays. The first eighth of these words in the frequency list (i.e. the most frequent 12.5%) take up 90.36% of the words altogether. In the Texts column you see 52.57% which means that the total number of plays over all the words are split up 52.57% in the first eighth only 2.38% in the most infrequent words.

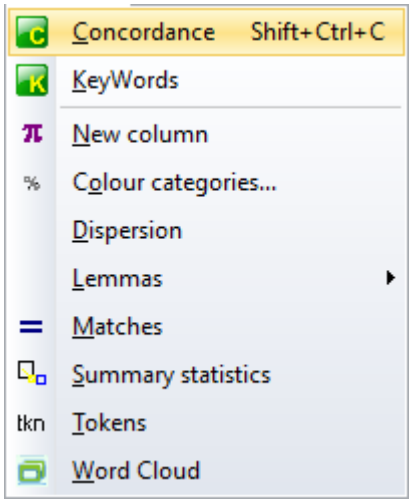
Eighths?

The divisions into 8ths are dependent on your plot ruler divisions 116.

See also: compute additional columns 65, add a Set column 33, save as Excel 91

5.7 compute concordance

With a key word list or a word list on your screen, you can choose *Compute* | *Concordance*



to call up a concordance of the currently selected word(s). The concordance will search for the same word(s) in the original text file that your list came from.

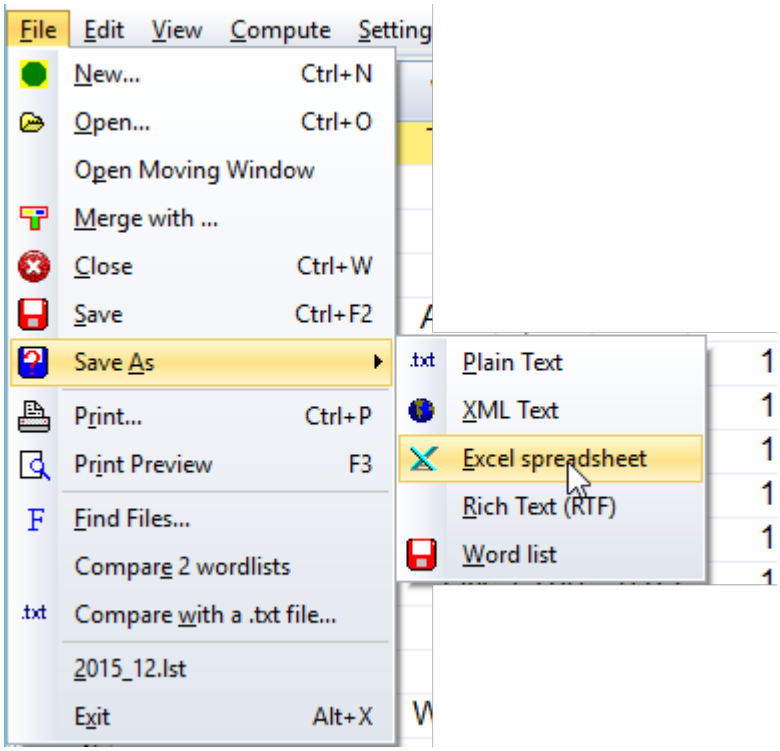
The point of it…

is to see these same words in their original contexts.

5.8 copy your results

The quickest and easiest method of copying your data e.g. into your word processor is to select with the cursor arrows and then press Ctrl+C. This puts it into the clipboard 529 (click to see examples showing how to copy into Word etc.).

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



saving as a text file or XML or spreadsheet 91

save 91 as data (not the same as saving as text: this is saving so you can access your data again another day)

See also: saving 91, printing 88, clipboard 529

5.9 editing

You can easily delete any item in a list by pressing *Del*. Press *Ins* to restore it.

If it's deleted it will look greyed out with a line through the wording.

6	IN	53,980	1.91	480	100.00	0.99
~~7~~	~~THAT~~	~~37,605~~	~~1.33~~	~~479~~	~~99.79~~	~~0.97~~
~~8~~	~~#~~	~~33,326~~	~~1.18~~	~~430~~	~~89.58~~	~~0.93~~
~~9~~	~~IS~~	~~32,211~~	~~1.14~~	~~480~~	~~100.00~~	~~0.99~~
10	IT	28,435	1.00	478	99.58	0.98

Deleted lines can be zapped 154 if you don't wish to see them any more, which removes them permanently.

The rest of this section concerns other ways of editing your data.

5.9.1 reduce data to n entries

With a very large word-list, concordance etc., you may wish to reduce it randomly (eg. for sampling). This menu option (*Edit* | *Deleting* | *Reduce to N*) allows you to specify how many entries you want to have in the list. If you reduce the data, entries will be randomly zapped 154 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 154, editing a list of data 75.

5.9.2 reverse-delete

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

See also : Delete or Restore to End 71.

5.9.3 delete or restore to end

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

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

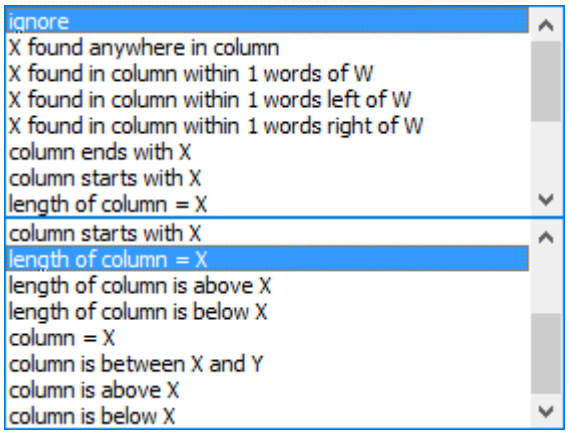
5.9.4 delete if

The idea is to be able to delete entries automatically using criteria you define. For example, a word list based on lots of texts but you want to remove any items which are from too few texts.

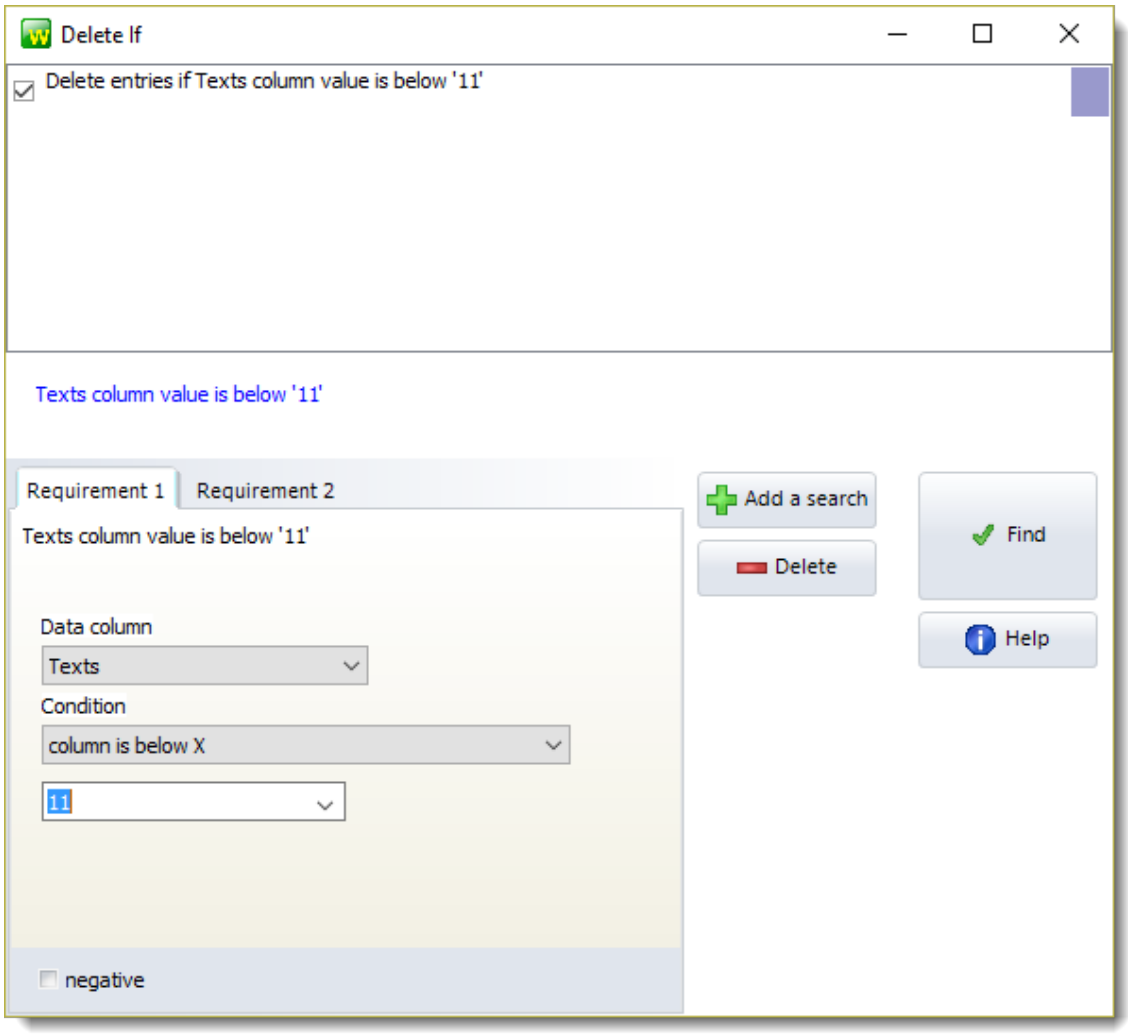
How to do it

Choose *Edit* | *Delete* | *Delete If*...

The data column box offers all the columns in your current data set. The conditions available include

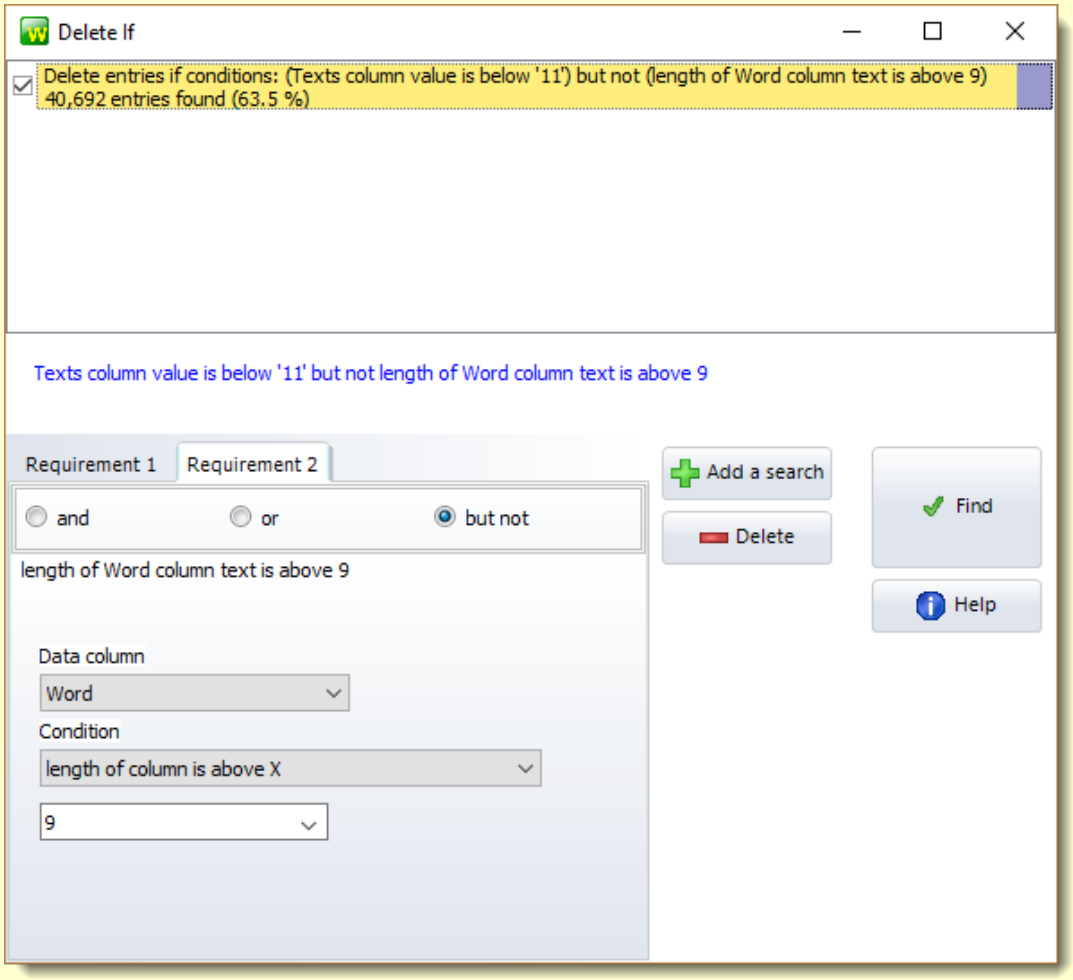


Simple Example



Advanced uses

If you check the negative box you will get the opposite effect. The list at the top shows you what effect to expect. You can choose several different options and there can be 2 different requirements, such as to delete all words which come from few texts unless they are long words.



In this example, we got this:

9	AARGH	3		3	0.63	0.00
10	AARHUS	2		2	0.42	0.35
11	AARON	3		3	0.63	0.51
12	AARSETH	1		1	0.21	
13	AART	1		1	0.21	
14	AAYAMUGUCHI	1		1	0.21	
15	AB	5		5	1.04	0.40
16	ABA	1		1	0.21	
17	ABABA	8		2	0.42	0.13
18	ABACK	8		8	1.67	0.58
19	ABACUS	3		3	0.63	0.51
20	ABALANCE	1		1	0.21	
21	ABALKIN	8		3	0.63	0.13
22	ABANAZER	1		1	0.21	
23	ABANDON	43		37	7.71	0.89
24	ABANDONDED	1		1	0.21	
25	ABANDONED	78		68	14.17	0.88

Once you've chosen the settings, press *Add a search* and *Find*. In the first example above all entries coming from 10 or fewer texts were deleted. If you decide that's OK, you can then Zap 154 them to get rid of them.

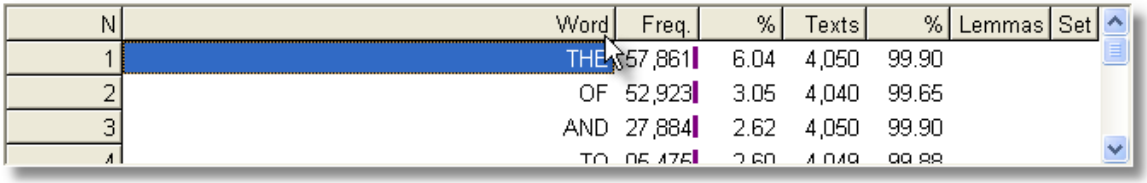
Changed your mind?

To restore any deleted items, go to the top of your list and choose *Edit* | *Restore deleted* | *Restore to end.*

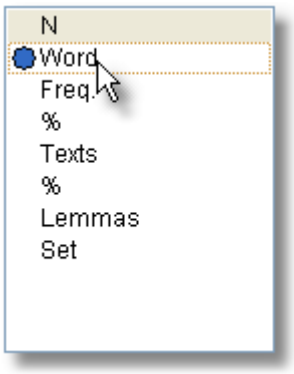
See also : colour categories 35

5.9.5 editing column headings

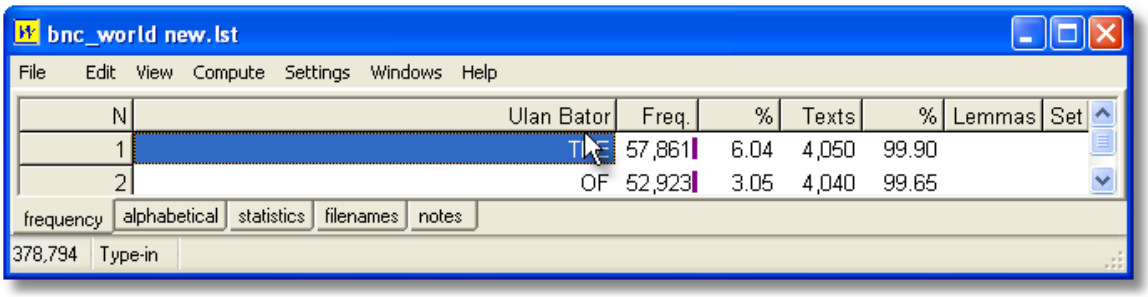
By default, a word-list will have column headings like these:



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



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



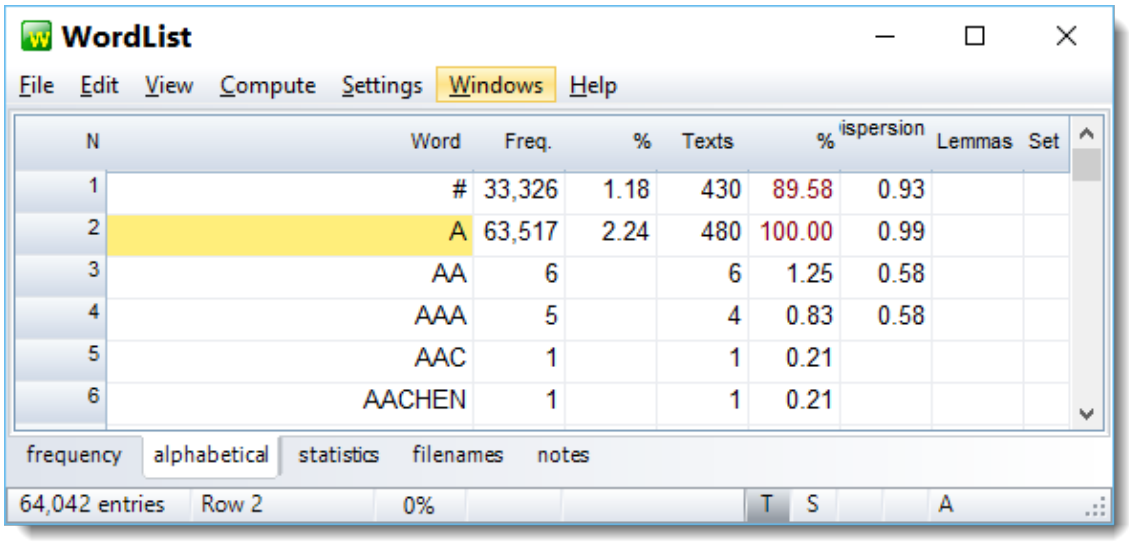
If you now save your word-list, the new column heading gets saved along with the data. Other new word-lists, though, will have the default WordSmith headings.

If you want *all future* word-lists to have the same headings, you should press the Save button in the layout window 60.

(If you had been silly enough to call the word column "Ulan Bator" and to have saved this for all subsequent word-lists, you could remedy the problem by deleting **`Documents\wsmith7\wordlist list customised.dat`**.)

5.9.6 editing a list of data

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



In the status bar at the bottom, the number in the first cell tells you how many words there are in the current word list, which row and which word (here **A**) you have selected.

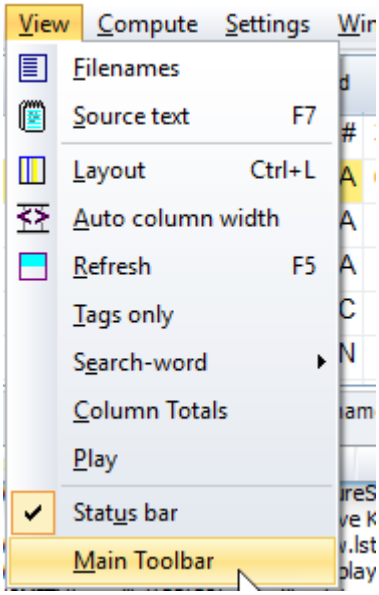
At the moment, when the user types anything, WordList will try to find what is typed in the list: the letter T is darkened.

If you click the S you will see it gets darkened, which changes the options for this list to *Set* (to classify your words, e.g. as adjectives v. nouns).

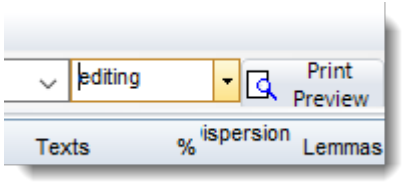
How to edit

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.

Open the main toolbar.



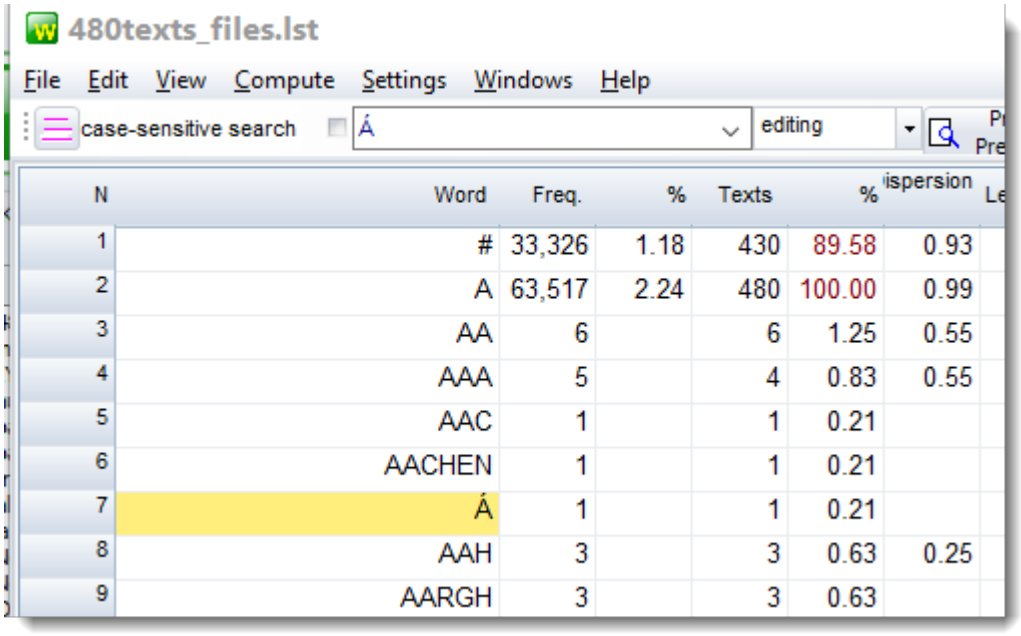
Choose editing in the main toolbar.



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

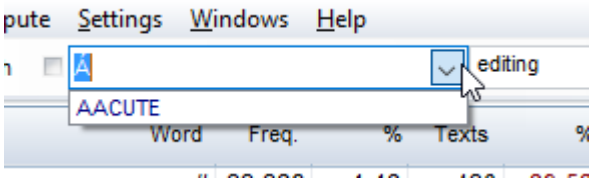
This will show the toolbar (if it wasn't visible before) so you can alter the form of the word or its frequency. If you spell the word so that it matches another existing word in the list, the list will be altered to reflect your changes.

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



N	Word	Freq.	%	Texts	%	ispersion	Le
1	#	33,326	1.18	430	89.58	0.93	
2	A	63,517	2.24	480	100.00	0.99	
3	AA	6		6	1.25	0.55	
4	AAA	5		4	0.83	0.55	
5	AAC	1		1	0.21		
6	AACHEN	1		1	0.21		
7	Á	1		1	0.21		
8	AAH	3		3	0.63	0.25	
9	AARGH	3		3	0.63		

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 [548] ( ), and if you have changed a word such as **AAAAAGH** to a pre-existing word such as **AAGH**, to join [322] the two entries.

See also: joining entries [322], finding source files [537].

5.10 language

WordSmith Tools handles a very wide range of languages, ranging from Albanian to Zulu. Chinese, Japanese, Arabic etc. are handled in Unicode.

The point of it …

1. Different languages sometimes require specific fonts.

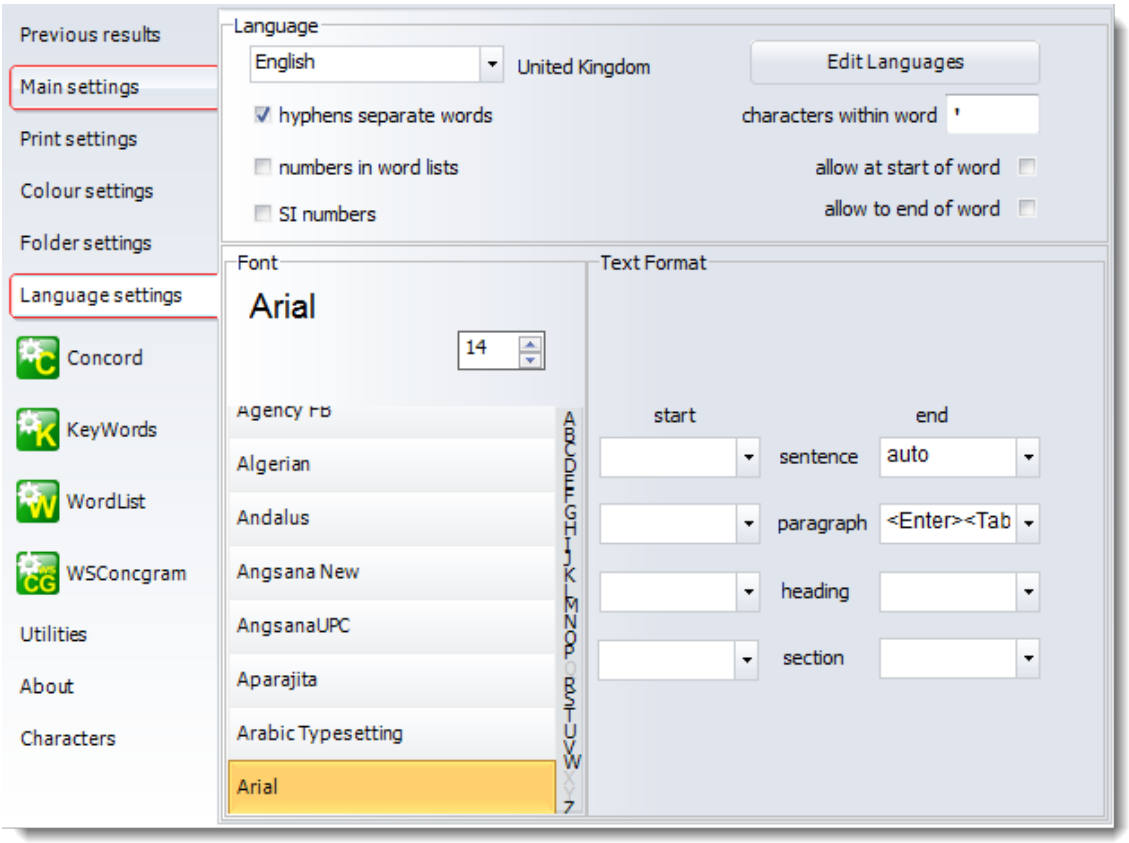
2. Languages vary considerably in their preferences regarding sorting order. Spanish, for example, uses this order: A,B,C,CH,D,E,F,G,H,I,J,K,L,LL,M,N,Ñ,O,P,Q,R,S,T,U,V,W,X,Y,Z. And accented characters are by default treated as equivalent to their unaccented counterparts in some languages (so, in French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

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

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

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

Settings



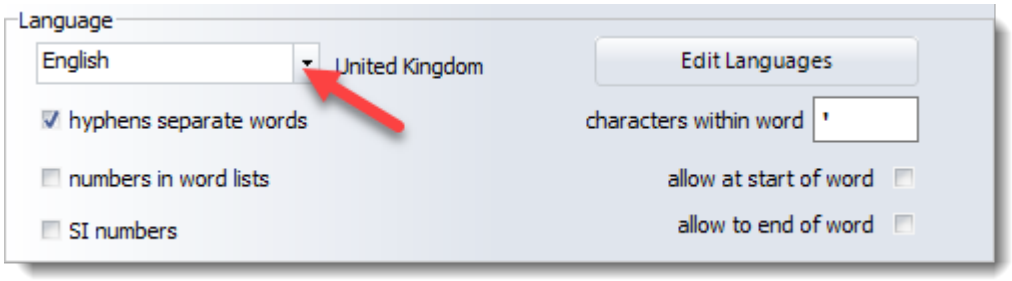
Hyphens separate words, Numbers, Characters within word 145

Font 118

Text Format 145

Not English?

Choose the language for the text you're analysing in the Controller 5 under *Language Settings*.



If the list which opens up where the red arrow shows doesn't include your language, press *Edit languages*.

The language and character set 527 must be compatible, e.g. English is compatible with Windows Western (1252), DOS Multilingual (850).

You can view word lists, concordances, etc. in different languages at the same time.

See also: Choosing Accents & Symbols 528, Accented characters 527, Processing text in Chinese etc., Text Format 164, Changing language of saved data 79

5.10.1 Change language within a Tool

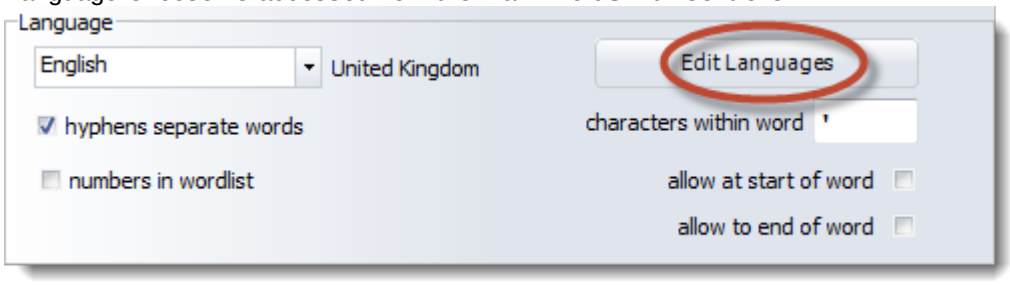
Most Tools use one language only and save information about that language which they are operating in, though Aligner can operate in lots of languages.

If you need to change language at some point, choose *Edit* | *Change language* ( ) in the menu.

5.10.2 Language Chooser

How to get here

The Language Chooser is accessed from the main WordSmith Controller:



or a button in the list of Utilities 8.

See also : Other Languages 82, saving your choices 81

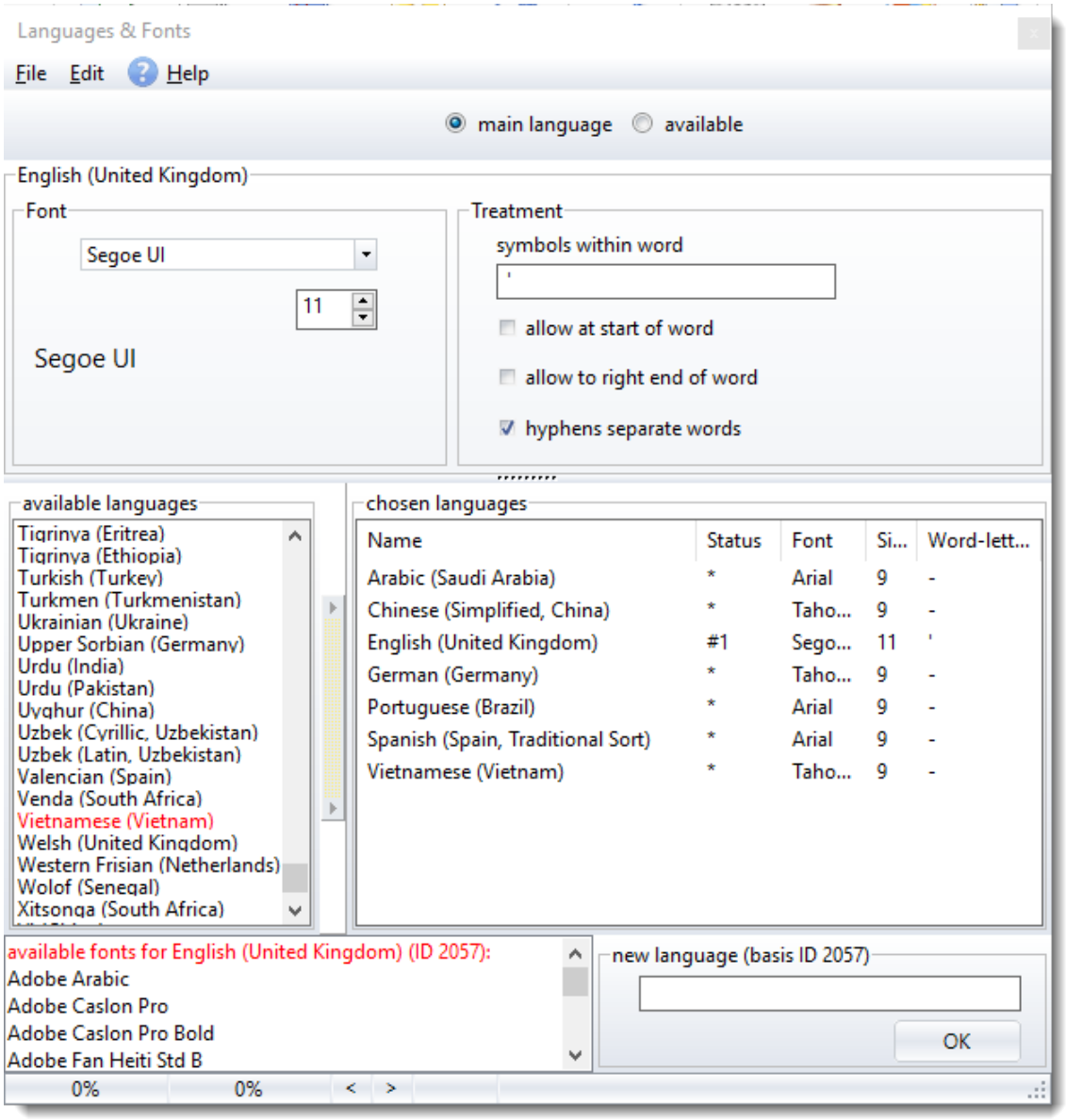
5.10.2.1 Overview

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

How to get here

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

What you will see may look like this:



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 *. Here, 7 languages have been chosen already.

To change the status of a chosen language, click the radio button at the top. To delete any unwanted language, right-click and choose "delete". To add a language, drag it from the left window to the right, then set the font and treatment you prefer for that particular language.

For each chosen language, you can specify any symbols which can be included within a word, e.g. the apostrophe in English 146, 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).

At the bottom you will see what the fonts on your system are for the current language selected. Each time you change language, the list of fonts 81 available changes.

Some languages do not mark word-separators 534.

See also : Language 79, Font 81, Other Languages 82, saving your choices 81, Changing the language of saved data 526

5.10.2.2 saving your choices

Save your results before quitting, so that next time WordSmith Tools will know your preferences regarding fonts and your #1 default language and your subsidiary default languages and you won't need to run this again. Results will be in **`Documents\wsmith7\WS7_language_choices.ini`**.

See also : Language 79, Font 81, Other Languages 82

5.10.2.3 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 79, Other Languages 82, saving your choices 81

5.10.2.4 Other Languages

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

See also : Language 79, Font 81, saving your choices 81

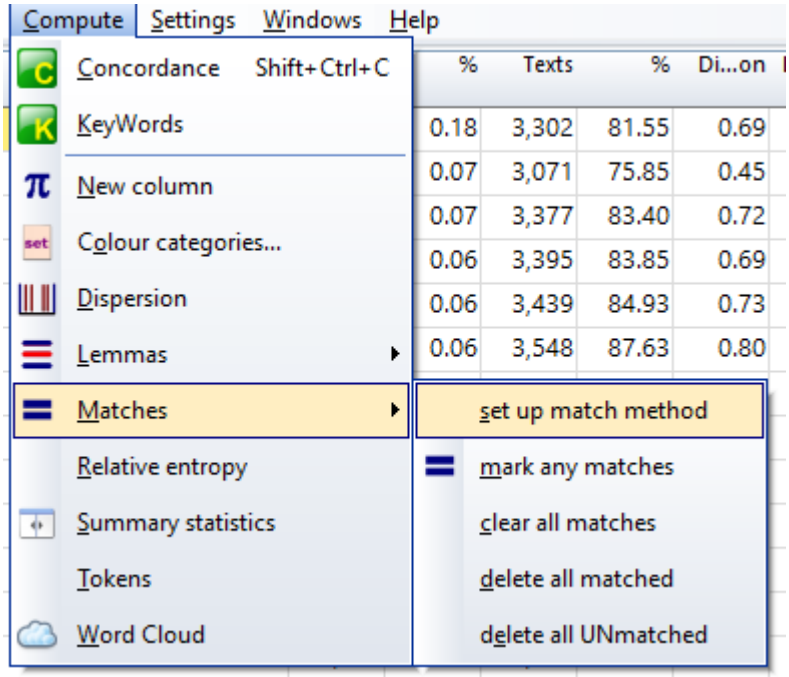
5.11 match words in list

The point of it…

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

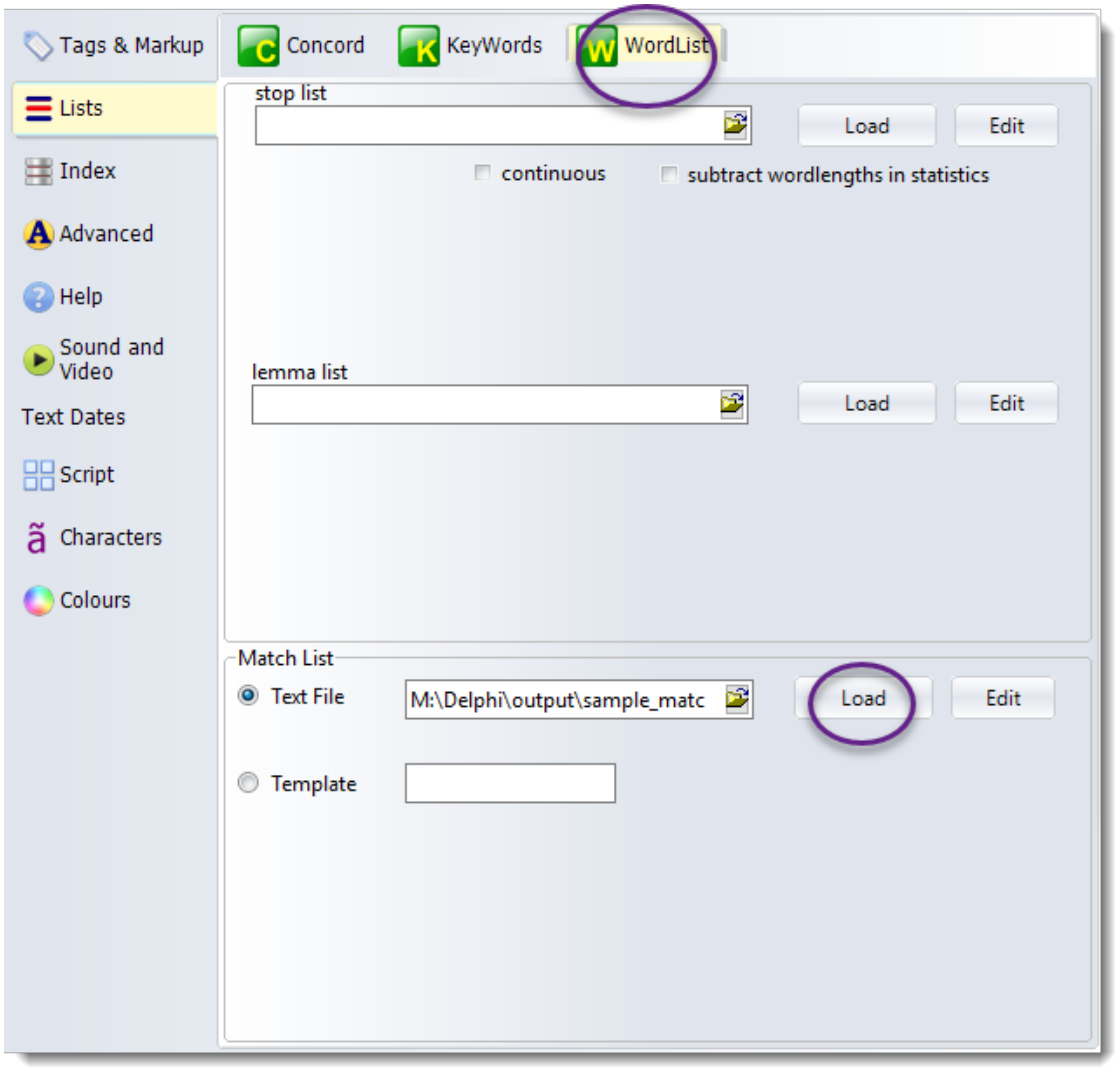
How to do it: WordList example

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



Set up match method

The main Controller settings dialogue box appears.



The circled areas show some of the main choices: make sure you are choosing for the right Tool, and if matching words from a text file, browse to find it and then press *Load* to load its words.

Text File or Template

You could 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 else a template filter such as ***ing** (which checks every entry to see whether it contains a word ending in **ing**.).

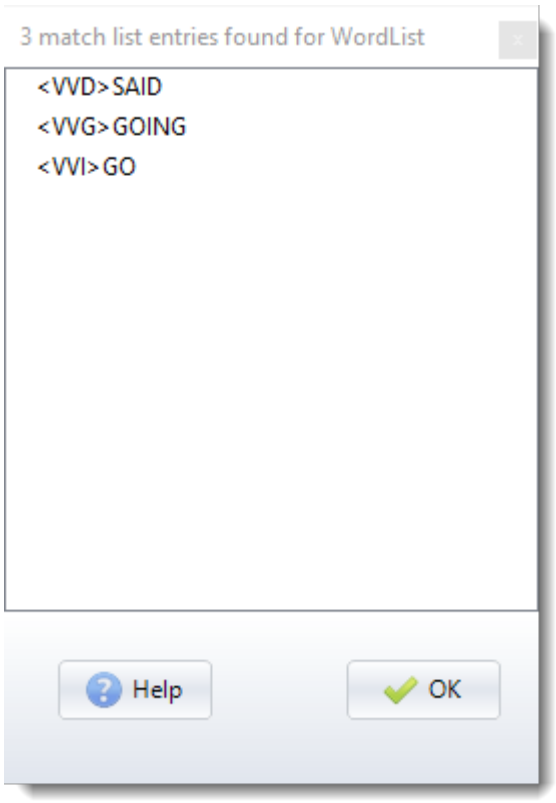
To use a match list in a file, you first prepare a file, using **Notepad** or any plain text word processor, which specifies all the words you wish to match up. Separate each word using commas, or else place each one on a new line. You can use capital letters or lower-case as you prefer. You can use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

Example

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

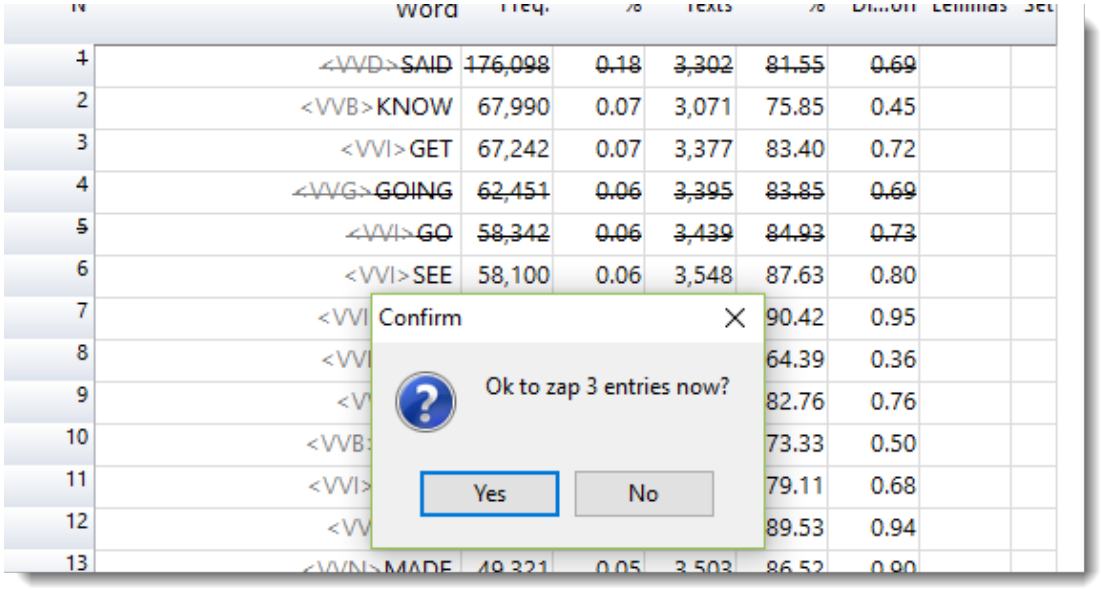
You may put angle-brackets in a stop list. So **<VERB>WANT** (that is **want** as a verb) will remove any cases of precisely that sequence if it is in your texts. It will only match **want** if marked with **<VERB>** in the corpus.

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

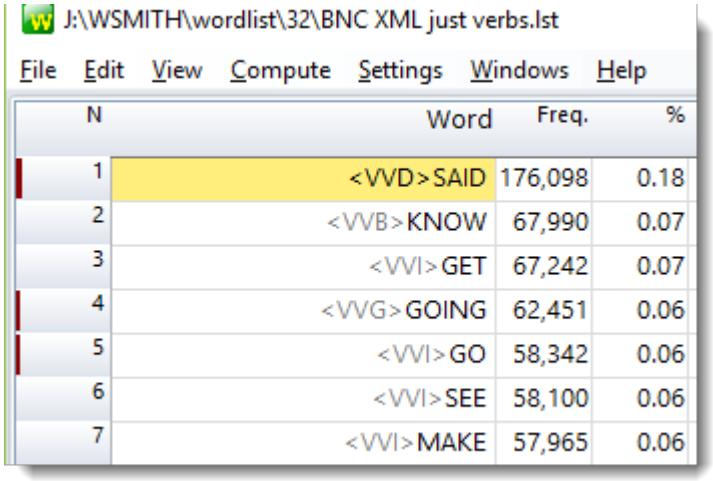


Action

If you have done the match settings the procedure 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 in the menu (as in the example below where 3 matching entries were found, the action selected was *delete all matched*.



In this screen-shot, the action was *mark any matches*. Matching entries are coloured in the left margin.



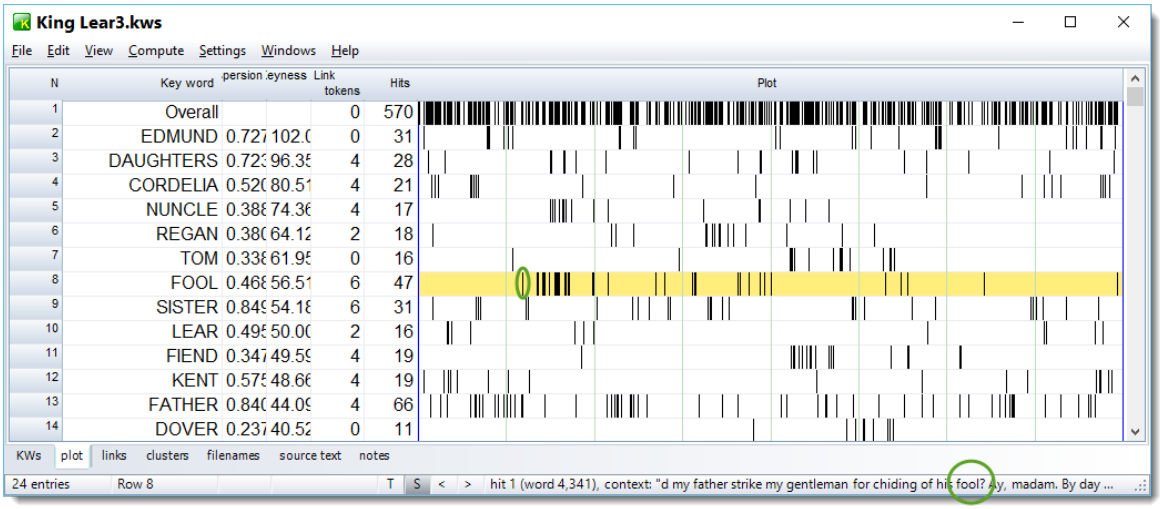
The marking can be cleared using the *clear all matches* menu option.

You can obtain statistics of the matches, using the Summary Statistics 141 menu option.

See also: Comparing Word-lists 308, Comparing Versions 311, Stop Lists 141, Lemma Matching 327, Text Converter Removing words 487

5.12 plot contexts

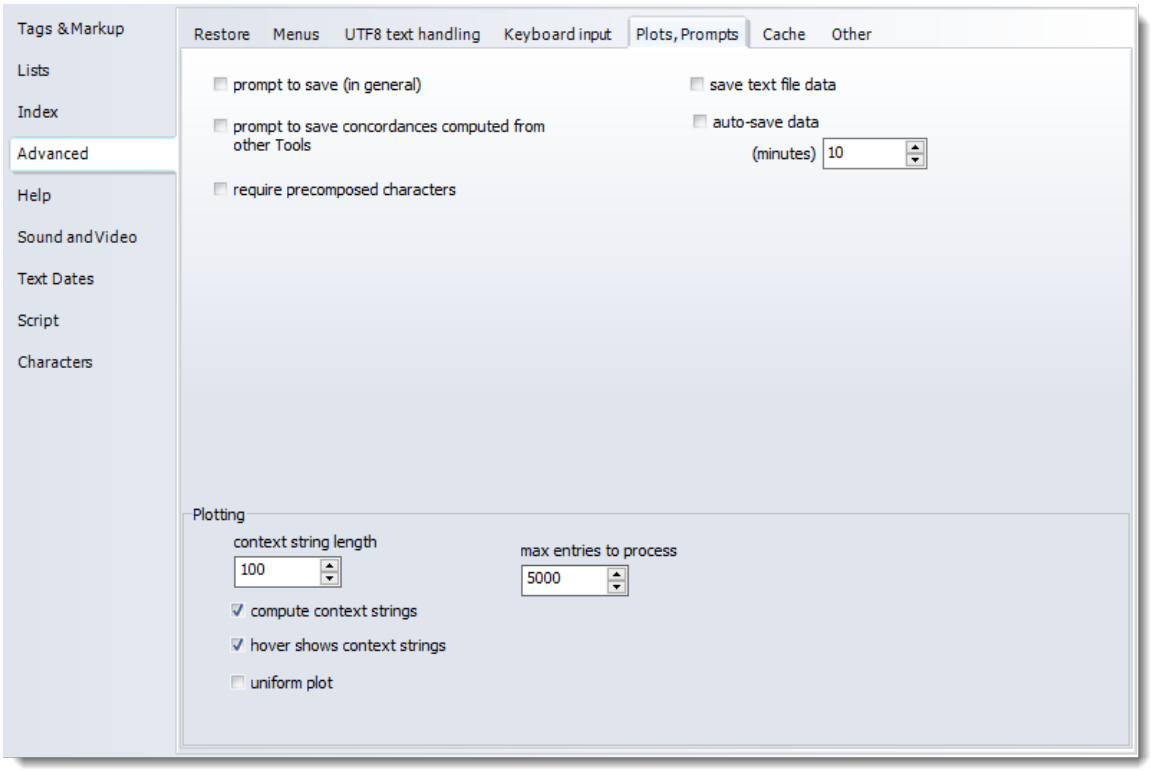
The plots as computed in the Concord and KeyWords tools can show the individual context of each mark.



In this screen shot the cursor was over the first vertical mark in the line referring to the key word **F00L**. In the status bar you see the context. The small < and > marks in the status bar let you go from one context to the next so as to see the detail of each plot mark.

Settings

Settings are found in the *Advanced Settings* | *Advanced* | *Plots,Prompts* section

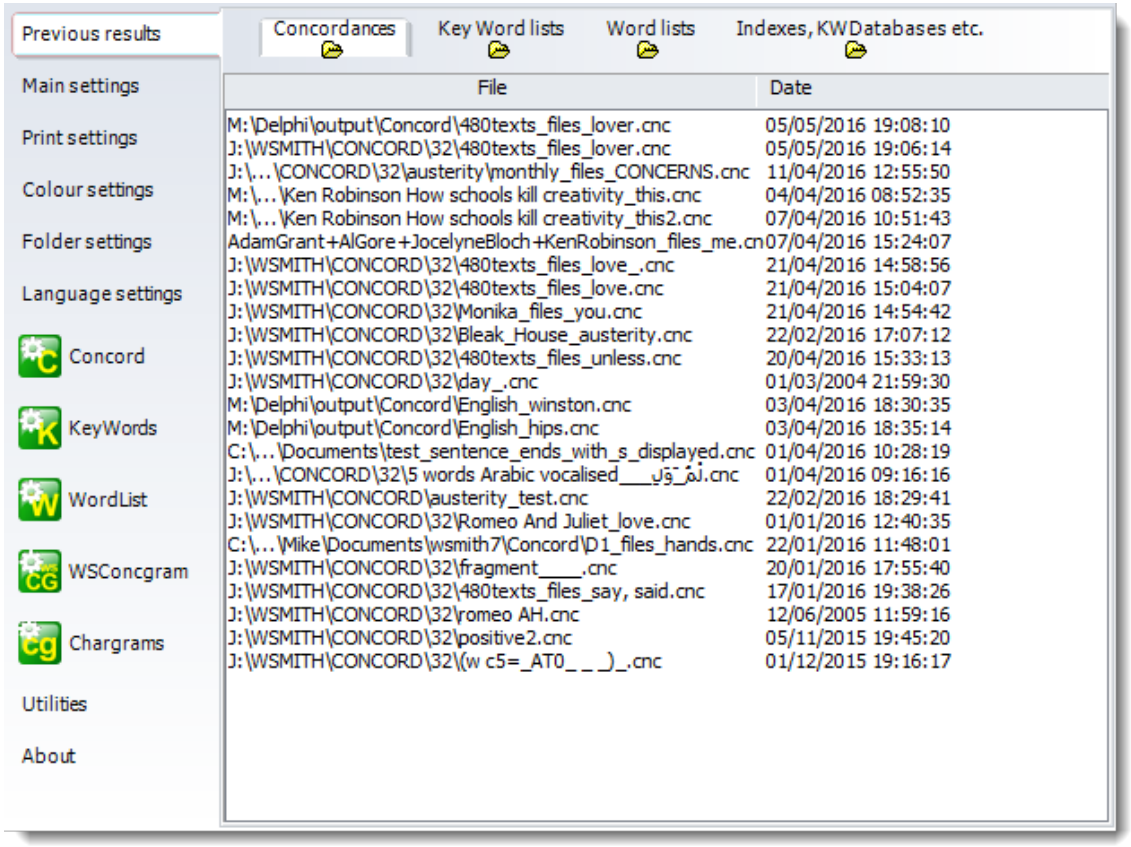


The amount of data saved with each plot can be considerable, so *max entries to process* allows you to save context strings only when there won't be too much data. A setting of 5,000 means that context strings only get included if there are no more than 5,000 concordance lines; in the case of key words no more than 5,000 key word contexts in the whole set of key words.

Hover shows context strings controls whether you see context strings when your mouse cursor hovers over plot marks. If un-checked, you can still see the contexts by pressing the < and > buttons in the status bar.

You can also double-click on a plot to see a list of all the plot entries 294.

5.13 previous lists



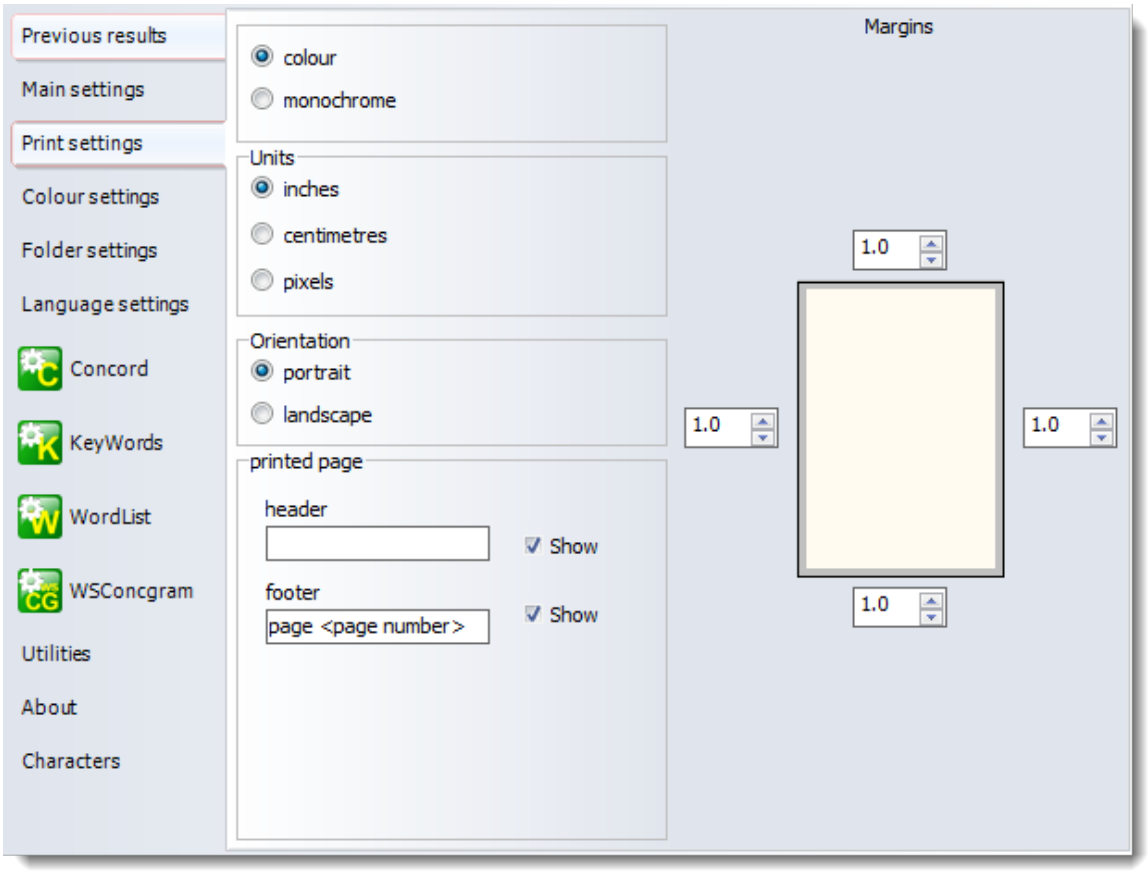
These windows show the lists of results you have obtained in previous uses of WordSmith.

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

To re-sort your entries click the header or choose *Resort* in the popup menu.

5.14 print and print preview

Print settings are in the main Controller:

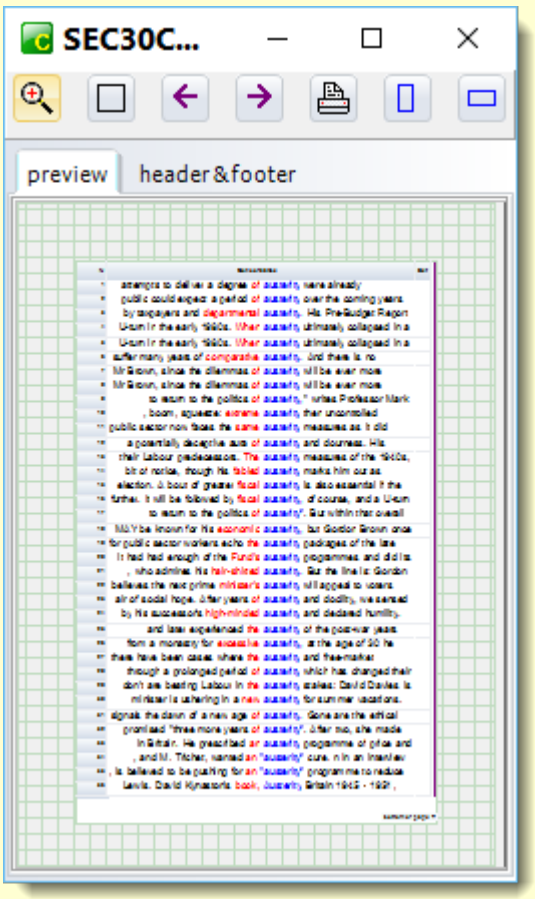


Print Settings

If you set printing to monochrome, your printer will use italics or bold type for any columns using other than the current "plain text" colour[114]. 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.

Print Preview

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



Bigger and Smaller

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

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

Next ( ) & Previous ( ) Page

Takes you forward or back a page.

Portrait ( ) or Landscape ( )?

Sets printing to the page shape you want.

Header, Footer, Margins

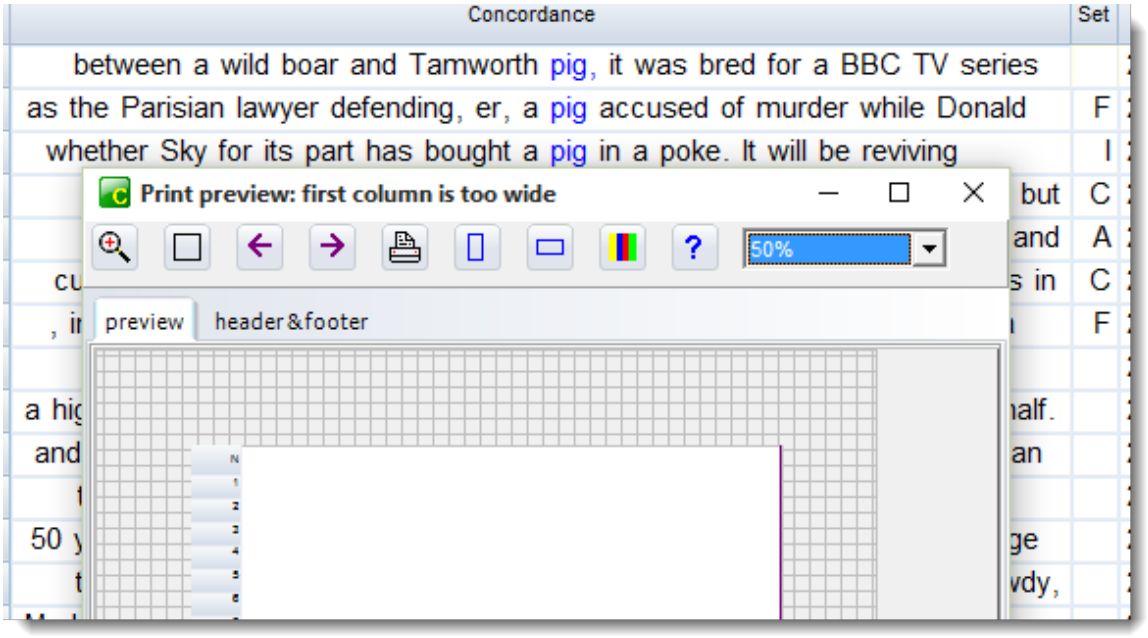
You can type a header & footer to appear on each page. Press *Show* if you want them included. If you include **<date>** this will put today's date and **<page number>** does the numbering. Margins are altered in the main Controller.

Print ( )

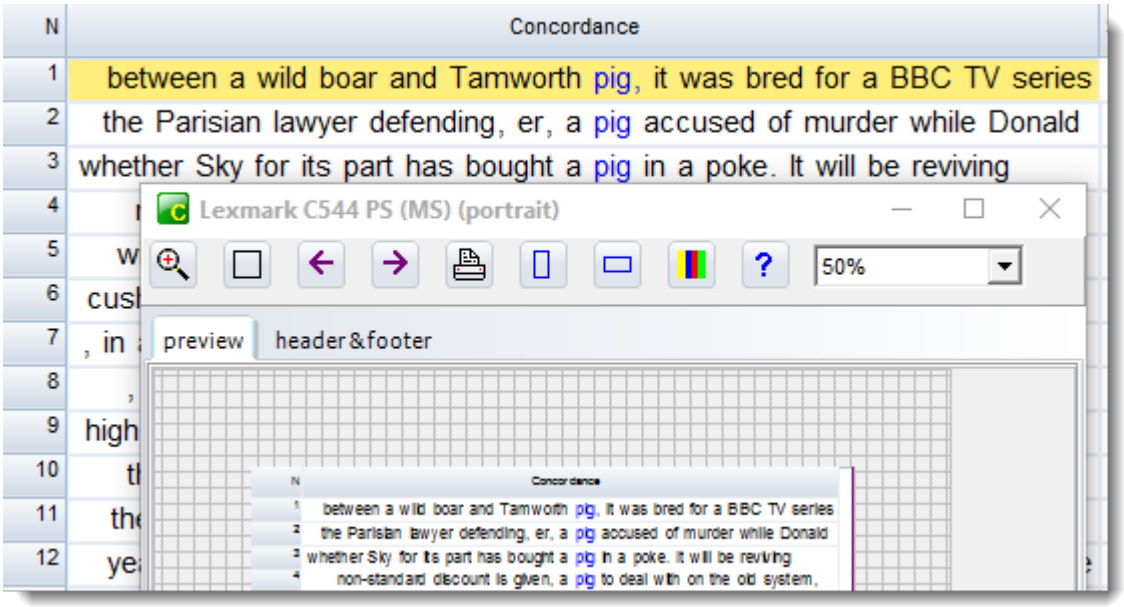
This calls up the standard Windows printer page and by default sets it to print the current page. You can choose other pages in this standard dialogue box if you want.

Some columns of data not shown

A case like this showing nothing but the line numbers



is because you have pulled the concordance data too wide for the paper. WordSmith prints only any columns of data which are going to fit. Shrink the column, hide 60 any unwanted ones, or else set the print surface to landscape.



See also: Printer Settings 113

5.15 quit WordSmith

Alt+X is the hot key.

Closing WordSmith Tools Controller 5 will close down all of the Tools.

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

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

By default, the last word list, concordance or key words listing that you saved or retrieved will be automatically restored on entry to WordSmith Tools. This feature can be turned off temporarily via a menu option or permanently in **`Documents\wsmith7\wordsmith7.ini`**.

5.16 saving

5.16.1 save results

To save your corrected results use *Save* (Ctrl+F2) in the menu. This saves all the results so you can return to the data at a later date. You may wish to clean up any deleted items by zapping 154, first.

Saved data is in a special **WordSmith Tools** format. The only point of it is to make it possible to use the data again another day. You will not be able to examine it usefully outside the Tools. If you want to export your data to a spreadsheet, graphics program, database or word processor, etc., you can do this either by saving as text 91 or by copying the data to the clipboard 529.

save As? ( ) 91

save part of the data only

By default, and save all your data that you haven't zapped 154. If you want to save only part of it, but don't want to zap it to oblivion, choose Copy 69.

avoid the save prompts?

You can avoid them in the Advanced settings 122 (*Main Settings* | *Advanced Settings* | *Advanced* | *Plots,Prompts*).

5.16.2 save as text, Excel etc.

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 529, e.g. if you simply want to insert your results into your word processor.

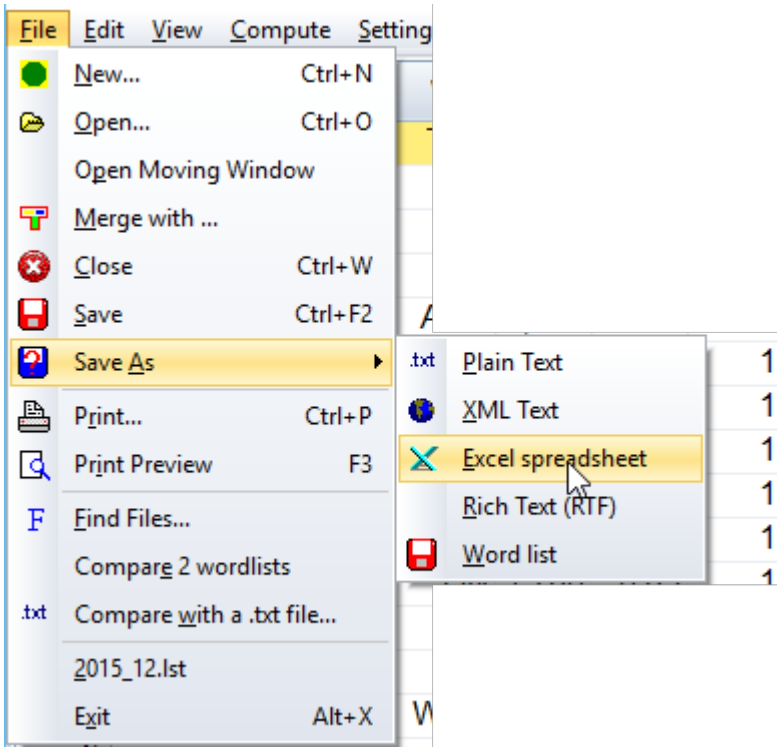
If you want to copy the data in colour, you should definitely use the clipboard 529.

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 as shown here or Print to File (via F3 or) or Copy ( ) to text file.

There are several options for saving in other formats



These menu choices up a Save as window as explained here 242 for Concord.

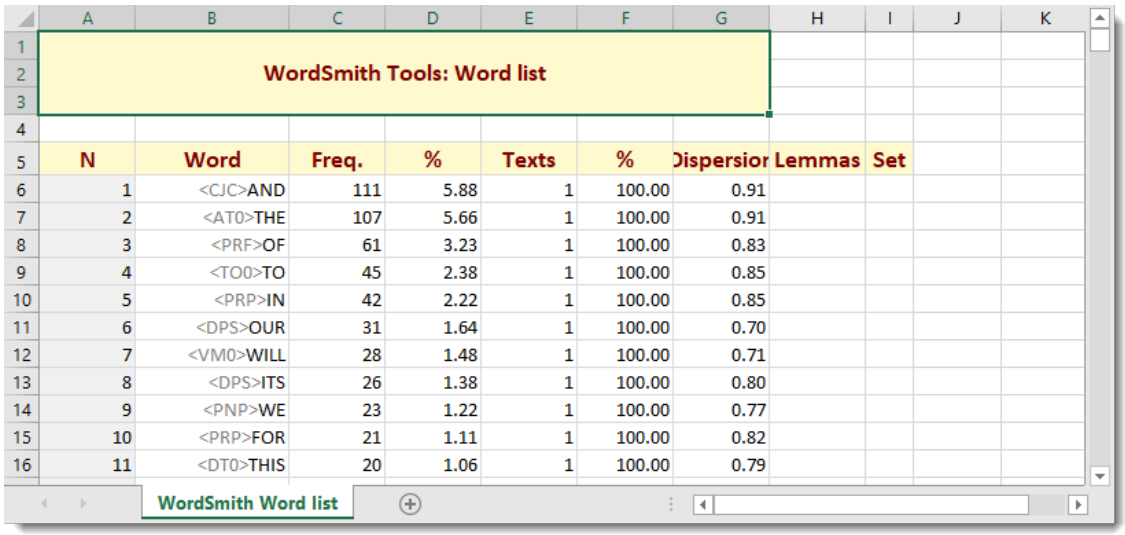
Options include:

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

columns	all/any which you have highlighted/a specific range (column 1 is the one with the numbers)

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

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

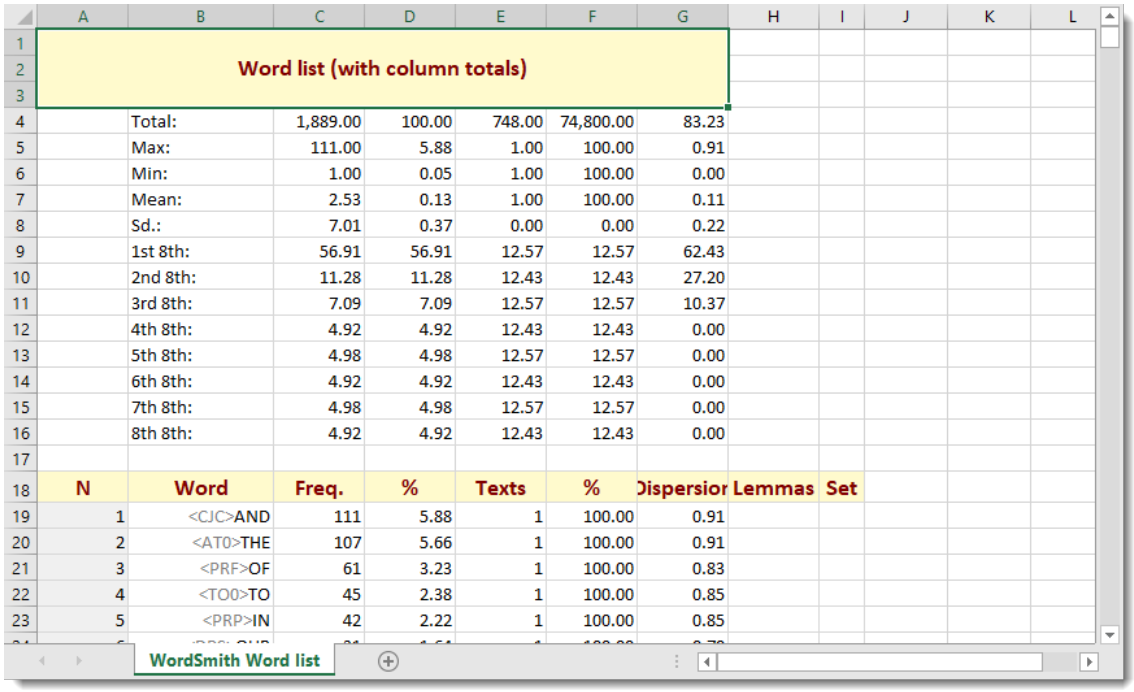


	A	B	C	D	E	F	G	H	I	J	K
1											
2				WordSmith Tools: Word list							
3											
4											
5	N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set		
6	1	<CJC>AND	111	5.88	1	100.00	0.91				
7	2	<AT0>THE	107	5.66	1	100.00	0.91				
8	3	<PRF>OF	61	3.23	1	100.00	0.83				
9	4	<TO0>TO	45	2.38	1	100.00	0.85				
10	5	<PRP>IN	42	2.22	1	100.00	0.85				
11	6	<DPS>OUR	31	1.64	1	100.00	0.70				
12	7	<VM0>WILL	28	1.48	1	100.00	0.71				
13	8	<DPS>ITS	26	1.38	1	100.00	0.80				
14	9	<PNP>WE	23	1.22	1	100.00	0.77				
15	10	<PRP>FOR	21	1.11	1	100.00	0.82				
16	11	<DT0>THIS	20	1.06	1	100.00	0.79				

WordSmith Word list

Note, here POS tags have been treated as prefixes 374.

or (with column totals 67):



	A	B	C	D	E	F	G	H	I	J	K	L
1												
2		Word list (with column totals)										
3												
4		Total:	1,889.00	100.00	748.00	74,800.00	83.23					
5		Max:	111.00	5.88	1.00	100.00	0.91					
6		Min:	1.00	0.05	1.00	100.00	0.00					
7		Mean:	2.53	0.13	1.00	100.00	0.11					
8		Sd.:	7.01	0.37	0.00	0.00	0.22					
9		1st 8th:	56.91	56.91	12.57	12.57	62.43					
10		2nd 8th:	11.28	11.28	12.43	12.43	27.20					
11		3rd 8th:	7.09	7.09	12.57	12.57	10.37					
12		4th 8th:	4.92	4.92	12.43	12.43	0.00					
13		5th 8th:	4.98	4.98	12.57	12.57	0.00					
14		6th 8th:	4.92	4.92	12.43	12.43	0.00					
15		7th 8th:	4.98	4.98	12.57	12.57	0.00					
16		8th 8th:	4.92	4.92	12.43	12.43	0.00					
17												
18	N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set			
19	1	<CJC>AND	111	5.88	1	100.00	0.91					
20	2	<AT0>THE	107	5.66	1	100.00	0.91					
21	3	<PRF>OF	61	3.23	1	100.00	0.83					
22	4	<TO0>TO	45	2.38	1	100.00	0.85					
23	5	<PRP>IN	42	2.22	1	100.00	0.85					

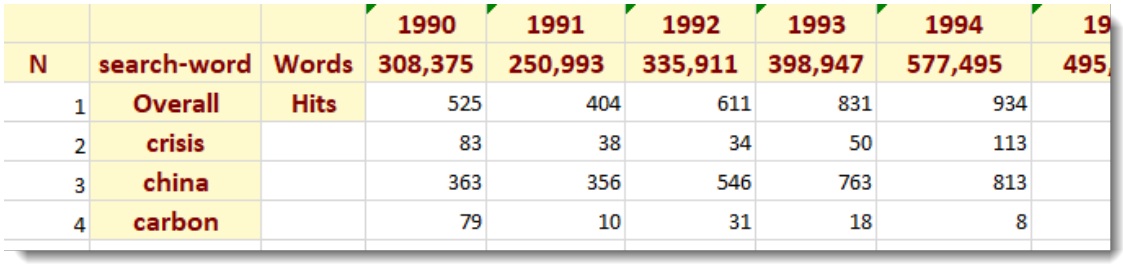
WordSmith Word list

The words are visible from row 19 onwards; above them we get some summary data. The 1/8, 2/8 etc. section splits the data into eighths; thus 57% of the Frequency data (column C) is in the 1st eighth.

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 253). The reason for this is that you will probably want a fixed number of characters, so that when using a non proportional font the search-words line up nicely. See also: Concord save and print 242.

Each worksheet can handle up to just over 1 million rows and over 16,000 columns. .

If your data contains a time-line 149 you will also get another worksheet in the Excel file, looking like this.



			1990	1991	1992	1993	1994	19
N	search-word	Words	308,375	250,993	335,911	398,947	577,495	495,
1	Overall	Hits	525	404	611	831	934	
2	crisis		83	38	34	50	113	
3	china		363	356	546	763	813	
4	carbon		79	10	31	18	8	

Details of a time-line 149 are found in a second tab in Excel, showing the word-counts for each time-period across the top and the number of hits for each search-word below.

The dispersion plot 230 of the same concordance data show in the second Excel tab, the positions of each hit for the various search-words in different texts. In the plot you see marks but here you get

the raw word numbers where each came. You could process those numbers in another program of your choice.

N	Positions										
1	16	17	19	21	24	24	26	29	29	30	30
2	239	263	278	619	1003	1014	1040	1096	1404	1440	1468
3	944	1025	1330	1677	1751	1883	2594	2608	2622	2737	3276
4	1113	1122	1130	1571	2060	2812	2852	2895	2930	3279	3600
5	43	120	131	200	267	318	345	364	394	429	464
6	33	90	211	295	439	510	564	578	636	675	778
7	35	97	104	121	286	441	494	534	571	608	773
8	12	175	252	342	394	571	573	701	729	787	791
9	72	322	652	1672	1870	2122	3426	3628	3733	4551	4809

In the case of XML text (X), you may get a large .XML file.

See also: Excel Files in batch processing 30, Automated scripts 123

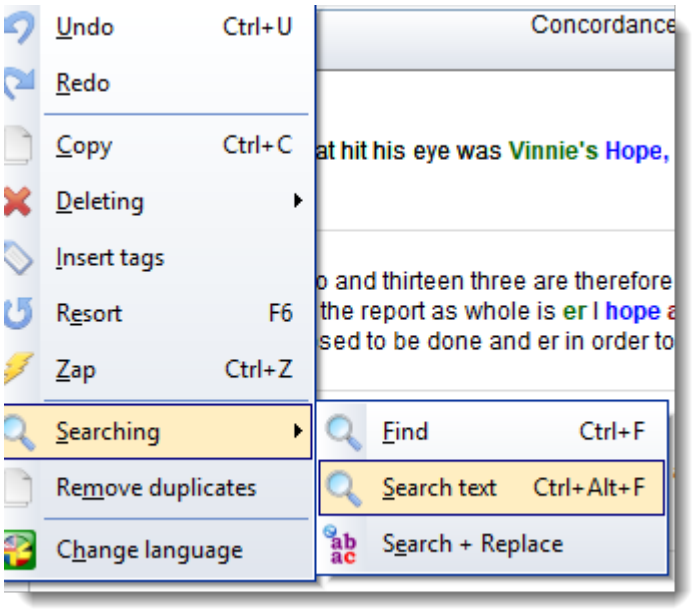
5.17 searching

5.17.1 three alternatives

All lists allow you to search for a word or part of one, or a number.

How to do it

Select the column of data you're interested in, then choose *Find, Search Text* or *Search + Replace* 102 in the *Edit* menu.



Find 96

Search Text 101

Search and Replace 102

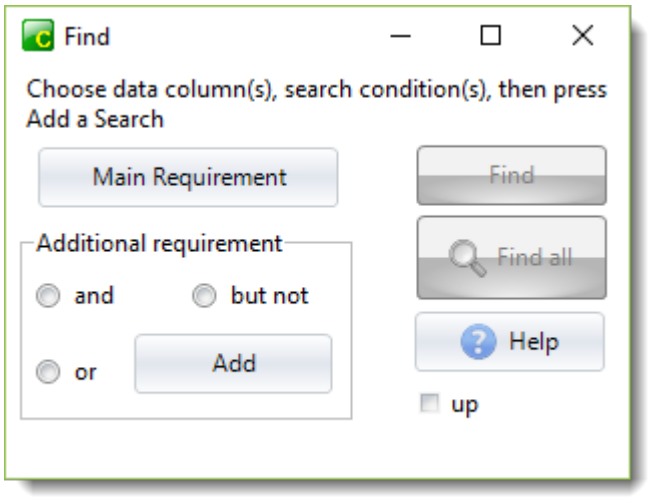
See also: Searching by Typing 106, Colour categories 35.

5.17.1.1 Find

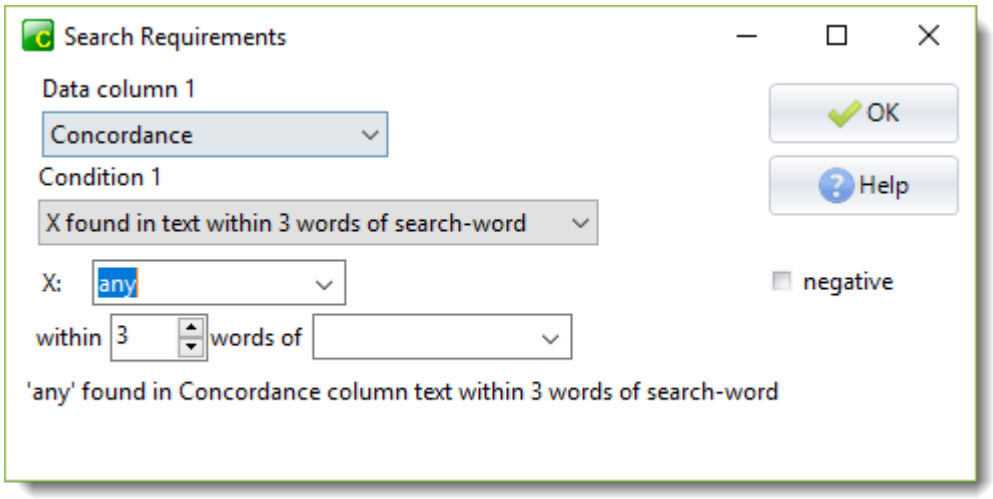
The Find window looks like this, a reduced form of the Colour Categories 99 window.

You can set two different criteria for searching.

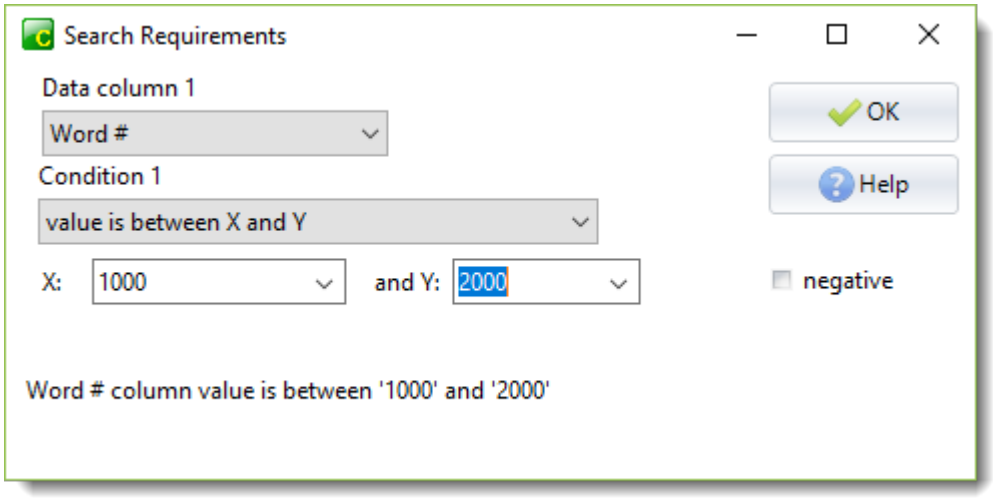
Press *Main Requirement* to tell WordSmith what you are looking for.



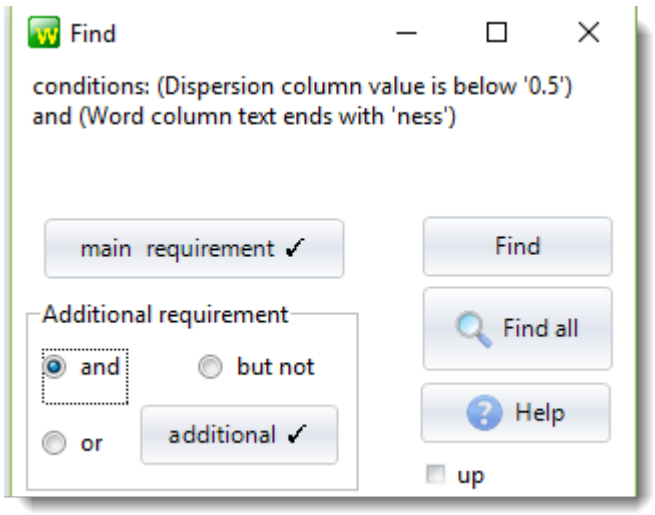
If the selected column is a Concordance column, you get options like these:



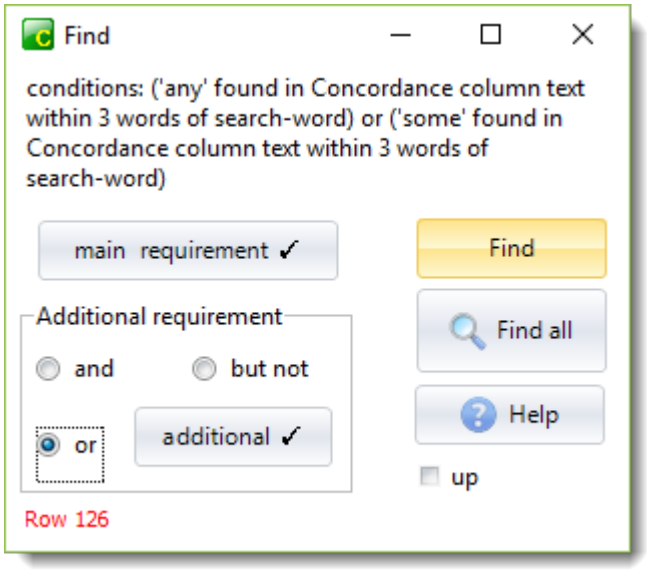
but if it's a numerical column you will see different options.



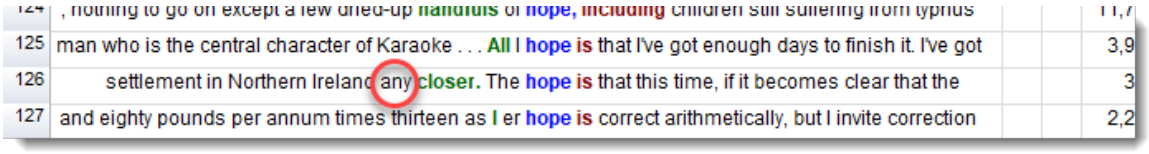
When you have chosen, press OK. You can add a further requirement if you like by choosing *and/but not/or* and pressing the *Additional* button:



Now press *Find* or *Find all.*

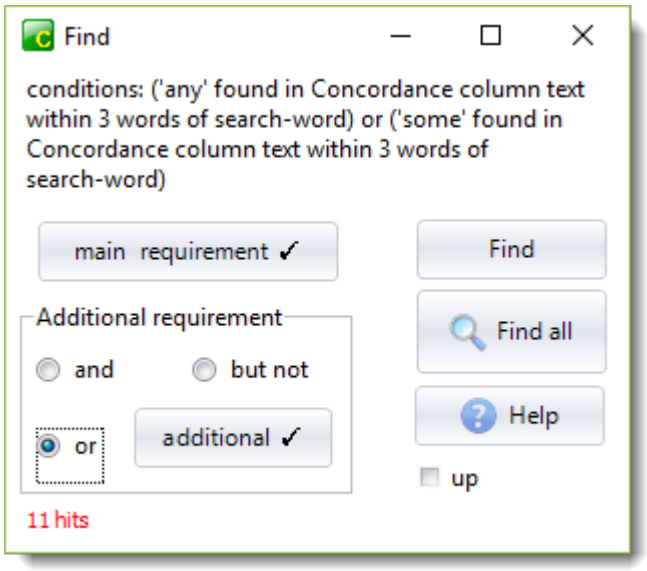


Here, *Find* was pressed, a hit was found at row 126 and the concordance was scrolled to it:



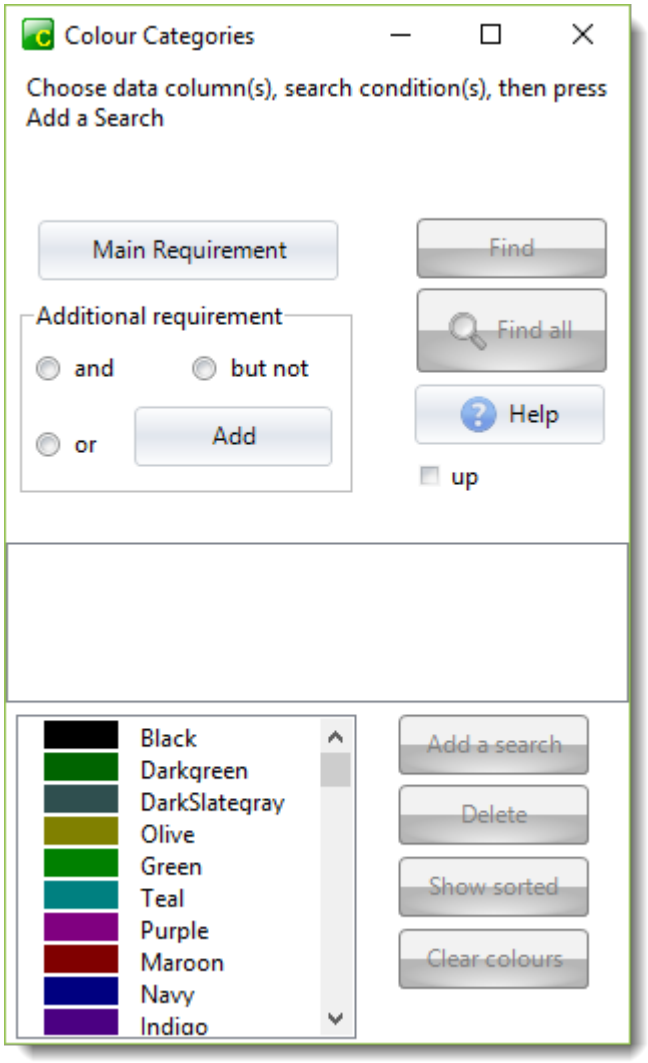
Find All?

This counts all rows where there is a hit and shows a total.



5.17.1.1.1 Colour categories find

The Colour Categories window looks like the Find 96 window but adds colour options.



Example

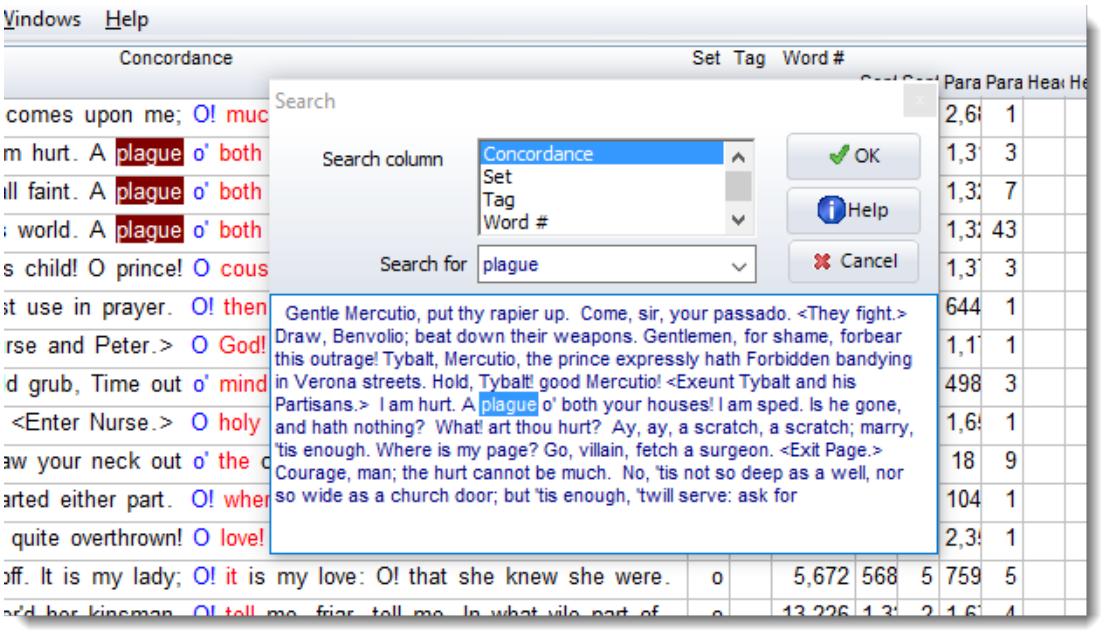
Here there were 2 requirements and 3 hits were found.

As in the Find 96 window, you can simply find and colour each one at a time by using the *Find* button.

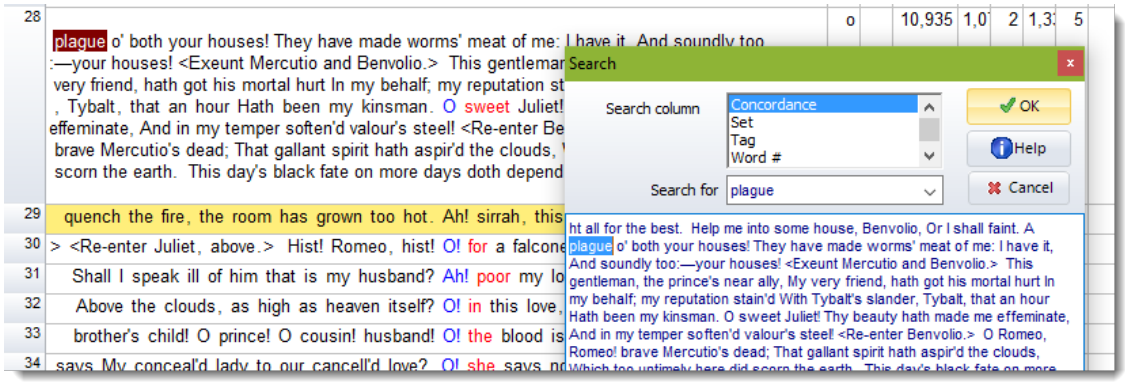
5.17.1.2 Search text

Search Text offers a search like this:

The search operates on the *current column* of data, though you can change the choice as in this screenshot, where Concordance is selected.

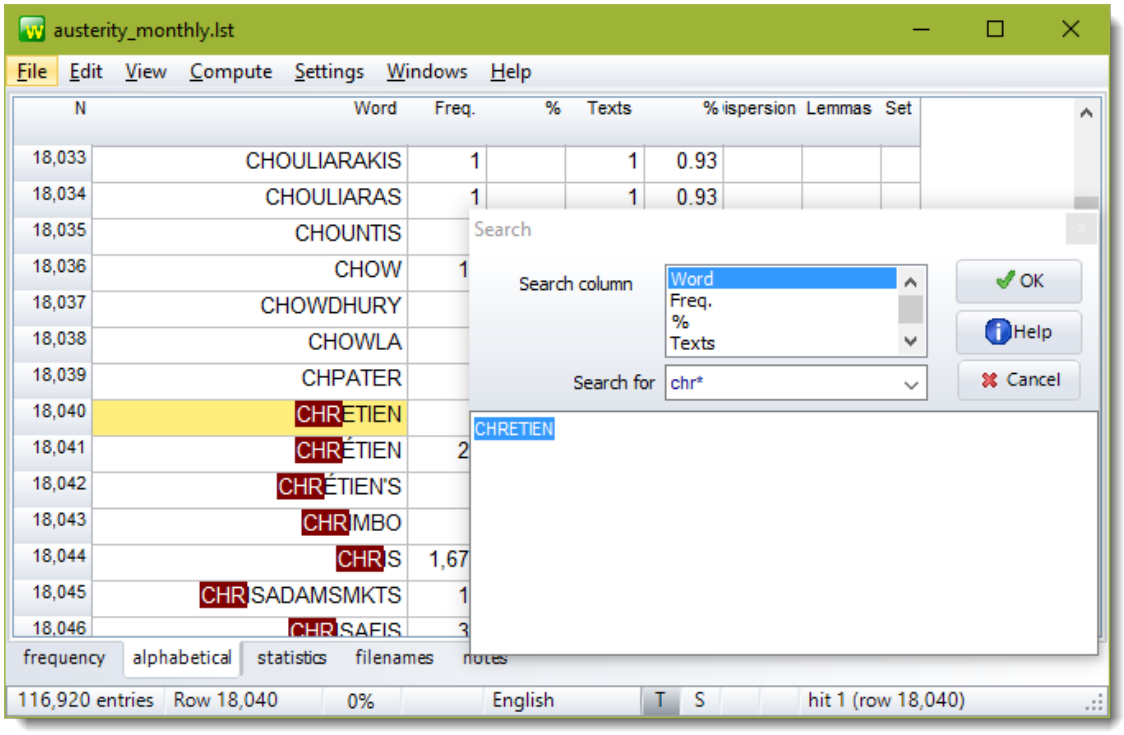


The syntax is as in Concord 246. In the case of a concordance line, the search operates over the whole context so far saved or retrieved. Here the word **plague** has been found near the blue concordance centre word but in line 28 it is far away:



To search again, press OK again....

In a word list you could get something like this:



Whole word – or bung in an asterisk

The syntax 246 is as in Concord, so by default a whole word search. To search for a suffix or prefix, use the asterisk as in the word list example here.

5.17.1.3 Search & replace

Some lists, such as lists of filenames 152, allow for searching and replacing.

The point of it

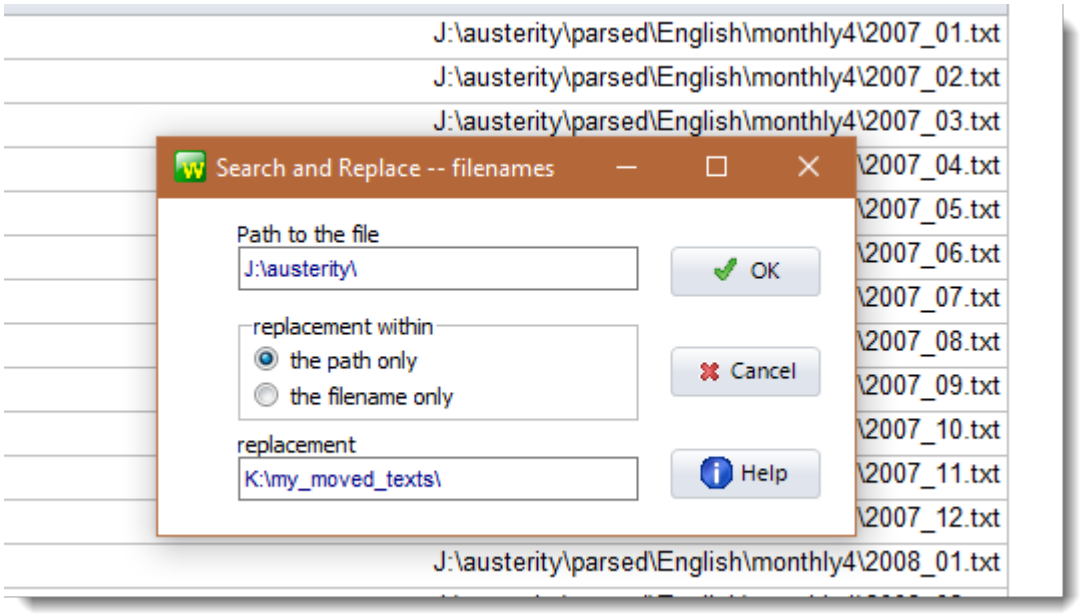
If your text data has been moved from one PC to another, or one drive to another, it will be necessary to edit all the filenames if WordSmith ever needs to get at the source texts, such as when computing a concordance from a word list 69.)

Search & Replace for file-names

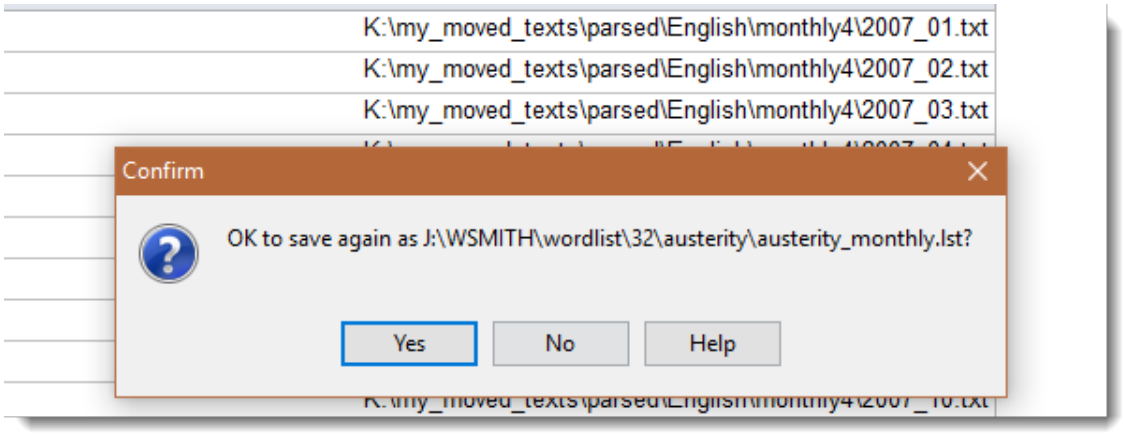
You may need to do this when WordSmith needs access to the original source texts as in a dispersion plot 230 or to show the source text 134 in full, if you have moved the text files.

If you are replacing a file-name you will see something like this. We distinguish between the path and the file's individual name, so that for a case like **`J:\austerity\parsed\English\monthly4\2007_01.txt`** the filename is **`2007_01.txt`** and the path to it is **`J:\austerity\parsed\English\monthly4`**.

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



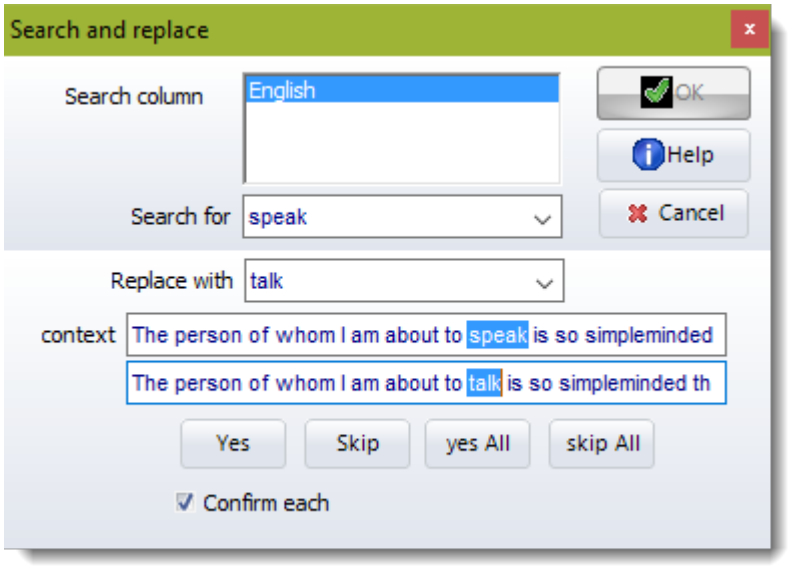
and all the filenames which contain **j:\austerity** will get **K:\my_moved_texts** as in this screenshot:



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

Search & Replace for other data

In this case the search & replace isn't of file-names but in the case below in *Aligner*, of the actual text. Like a search 95 operation, the search operates on the *current column* of data.



The context line shows what has been found.

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

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

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

Yes All: change without any check

Skip All: stop searching...

Whole word – or bung in an asterisk

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

Word lists can be sorted by suffix: see WordList sorting 372.

See also: Searching by Typing 106, Searching with F12 95, Accented Characters & Symbols 528.

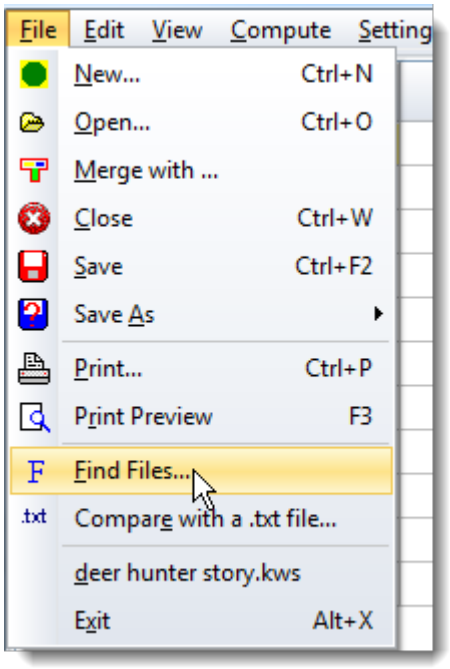
5.17.2 find relevant files

The point of it...

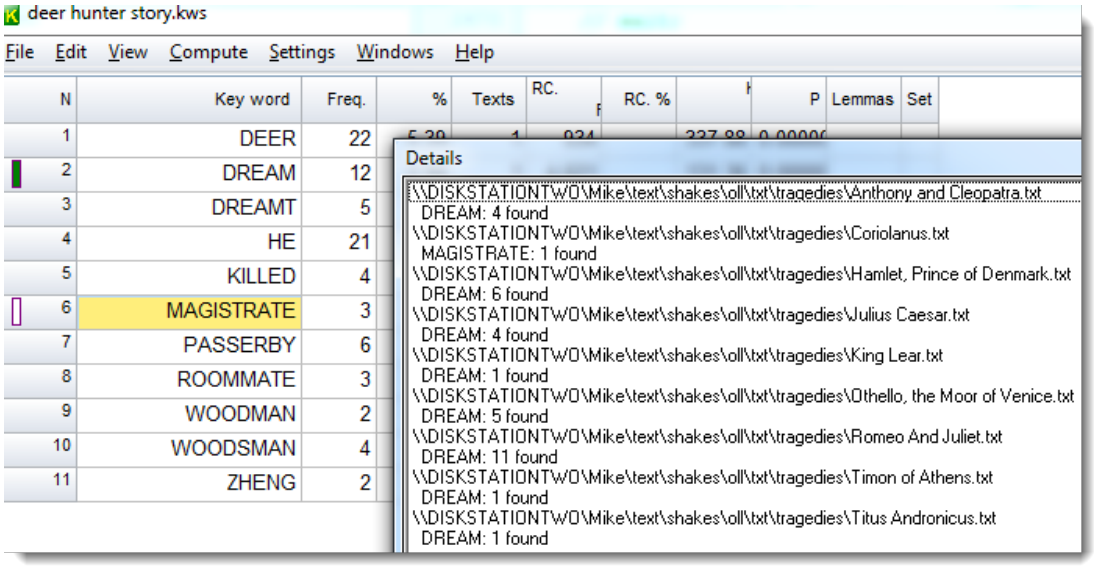
Suppose you have identified *muscle, fibre, protein* as key words in a specific text. You might want to find out whether there are any more texts in your corpus which use these words.

How to do it

This function can be reached in any window of data which contains the menu option, e.g. a word-list or a key words 7 list.



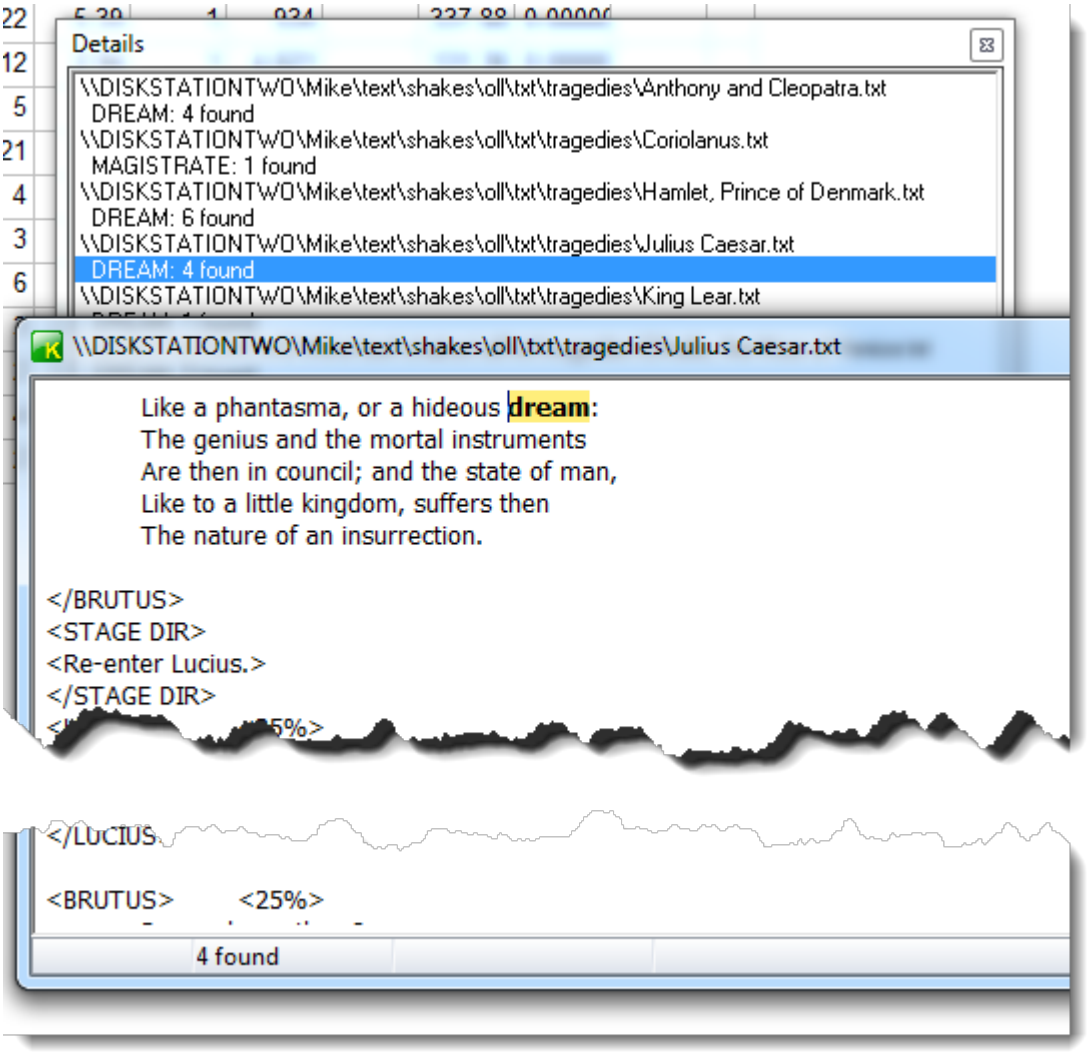
It enables you to seek out all text files which contain **at least one** mention of the words you have marked or selected. Before you click, choose the set of texts 52 which you want to peruse. (If you haven't, the function will let you use the text(s) the current key words or word-list entries are based on.)



Here we have a keywords list based on a Chinese folk tale with two items chosen by marking 108. The text files to examine in this case are all the Shakespeare tragedies...

What you get

A display based on all the words you marked, showing which text files 537 they were found in and how many of each word were found. If you double-click as shown

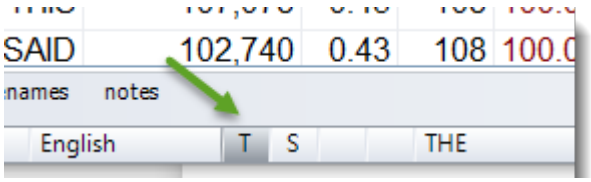


you'll get to see the source text and can examine each of the words, in this case the four tokens of the type *dream*.

See also: searching 95.

5.17.3 search by typing

The small T (type in) and S (Set) buttons in the status bar:



let you choose whether typing helps to quickly find a word by typing or changing a set category 43. As you type, **WordSmith** will get nearer.

See also: Edit v. Type-in mode 535, Searching for a word or part of one 95, Search & Replace 102, Editing 75, WordList sorting 372

5.18 selecting or marking entries

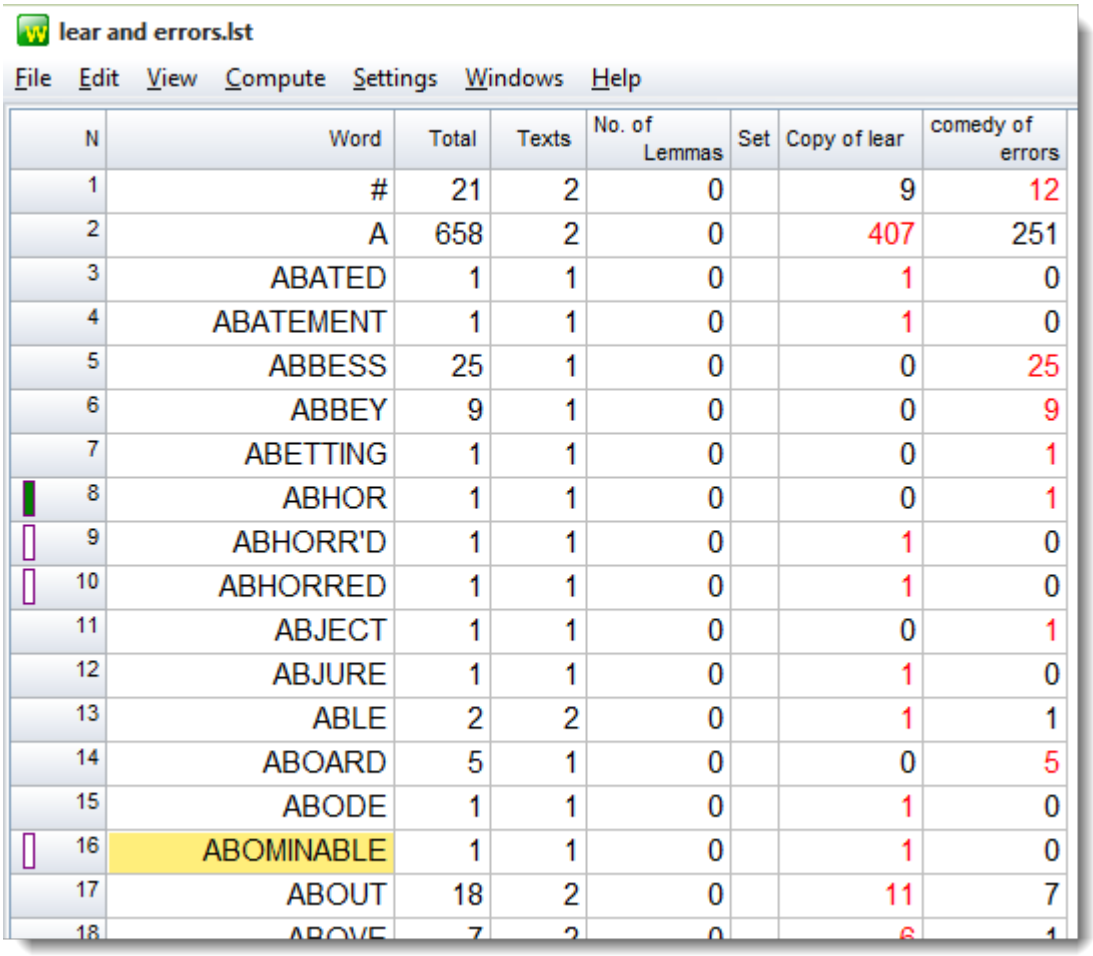
5.18.1 selecting multiple entries

It can be necessary to select non-adjacent entries e.g. for Joining 322 (lemmatisation).

How to do it

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

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

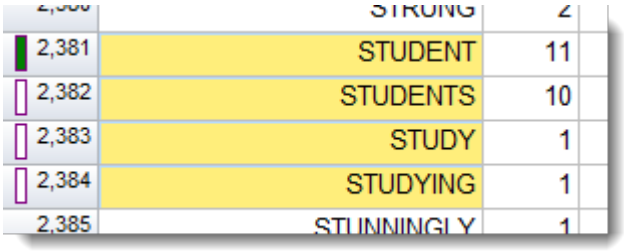


N	Word	Total	Texts	No. of Lemmas	Set	Copy of lear	comedy of errors
1	#	21	2	0		9	12
2	A	658	2	0		407	251
3	ABATED	1	1	0		1	0
4	ABATEMENT	1	1	0		1	0
5	ABBESS	25	1	0		0	25
6	ABBEY	9	1	0		0	9
7	ABETTING	1	1	0		0	1
8	ABHOR	1	1	0		0	1
9	ABHORR'D	1	1	0		1	0
10	ABHORRED	1	1	0		1	0
11	ABJECT	1	1	0		0	1
12	ABJURE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

Alternatively, to mark entries you can choose *Edit* | Mark (F5) 108 in the menu.

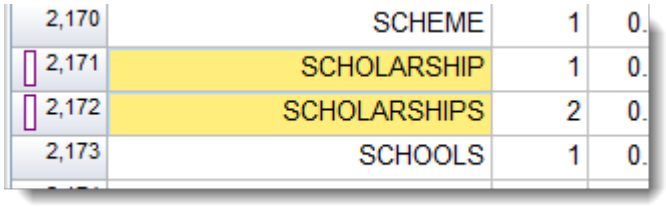
5.18.2 marking entries

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



2,381	STUDENT	11	
2,382	STUDENTS	10	
2,383	STUDY	1	
2,384	STUDYING	1	
2,385	STUNNINGLY	1	

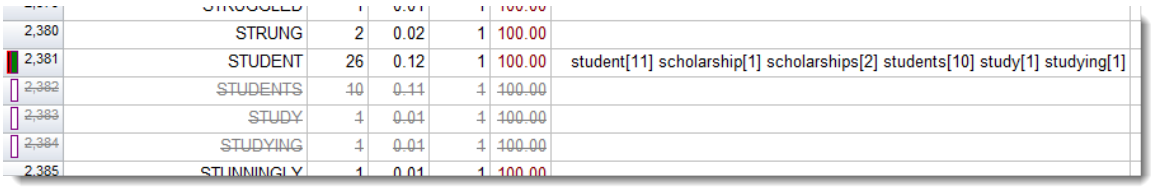
After marking these, I added **`scholarship(s)`** by selecting two more words then pressing Alt+F5 again.



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

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

To lemmatise all marked entries, press F4 after marking.



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

5.19 settings

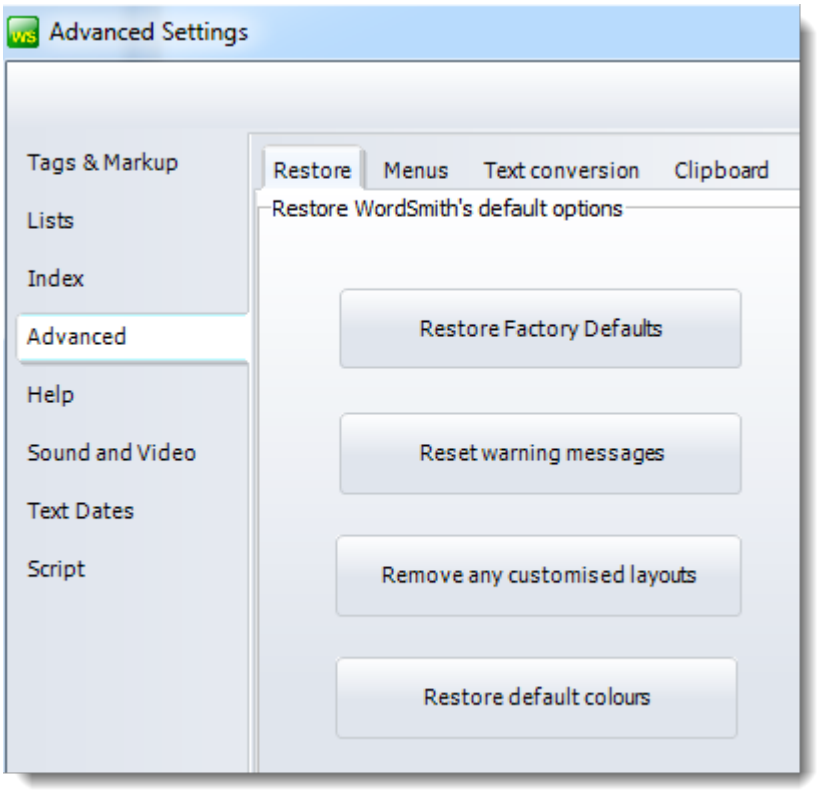
5.19.1 restoring settings

How to find it

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

The point of it...

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



Factory Defaults

Restores your **wordsmith7.ini** file to factory condition. Re-starts WordSmith with all relevant boxes filled in accordingly.

Warning Messages

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

Customised Layouts

Removes any of the layouts [60] you have saved for the various type of data WordSmith shows.

Colours

Re-sets colours [114] to the factory defaults.

5.19.2 save defaults

Settings can be altered by choosing *Colour Settings* in the WordSmith Tools Controller 5.
Any setting menu item in any Tool gives you access to these:

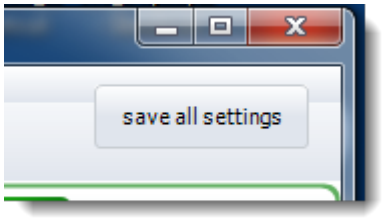
General, Folders, Colours, Languages, Tags, Lists, Concord, KeyWords, WordList, Index, Advanced, WSConcgram

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

colours 114	customise the default colours
folders 539	set WordSmith so it "knows" which folders you usually use
languages 145	character set 527, treatment of hyphens 544 & numbers, default file extension
general	restore last file 555, printing 113
tags 158	tags to ignore, tag file, tag file autoloading, custom tagsets 165
stop lists 141	for Concord, KeyWords and Wordlist
matching	files 82 to match up, or lemma files 322 to mark lemmas in a word list, etc.
Concord	number of entries, sort system, collocation horizons 207
KeyWords	procedure 287, max. p value 273, database & associate minimum frequencies, reference corpus 555 filename
WordList	word length & frequencies, type/token # 353, cluster 556 settings
Index 330	making a word-list index
Advanced 120	advanced settings
WSConcgram	for the concgram 13 utility

permanent settings and wordsmith7.ini file

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



Or by editing the **`wordsmith7.ini`** file, installed when you installed WordSmith Tools. This specifies all the settings which you regularly use for all the suite of programs, such as your text and results folders 539, screen colours 114, fonts 118, the default columns 60 to be shown in a concordance, etc.

You can restore the defaults 108.

show help file

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

sayings

Using Notepad, you can edit **Documents\wsmith7\sayings.txt**, which holds sayings that appear in the main Controller 5 window, if you don't like the sayings or want to add some more.

site licence and CD-ROM defaults

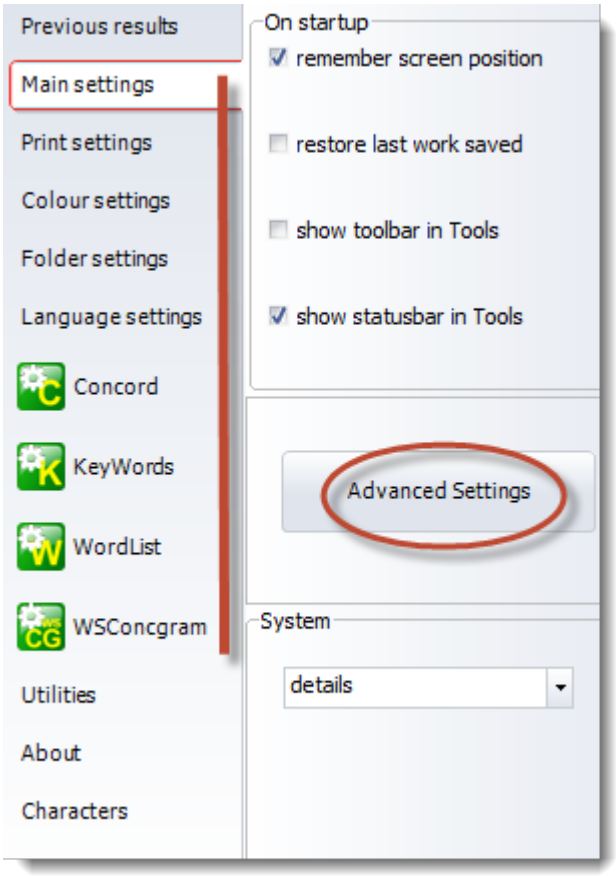
If you're running WordSmith straight from a CD-ROM, your defaults cannot be saved on it as it's read-only; Windows will find a suitable place for **wordsmith7.ini**, usually off the root folder of **My Documents**.

The first time you use WordSmith, you will be prompted to choose appropriate Folders 539, Text 145 Characteristics, Tag 158 details etc. and *Save All Settings* for future use. You can change settings and save them as often as you like.

Similarly, on a network you will usually not be allowed to change defaults permanently, as this would affect other users. Your network administrator should have installed the program so that you have your own copy of **wordsmith7.ini**, where it may be both read and altered. If WordSmith Tools finds a copy of **wordsmith7.ini** in that folder it will be able to use your personal preferences.

5.19.3 adjust settings

There are a number of Settings windows in the Controller 5. You will see tabs accessing them at the left in he main Controller window.

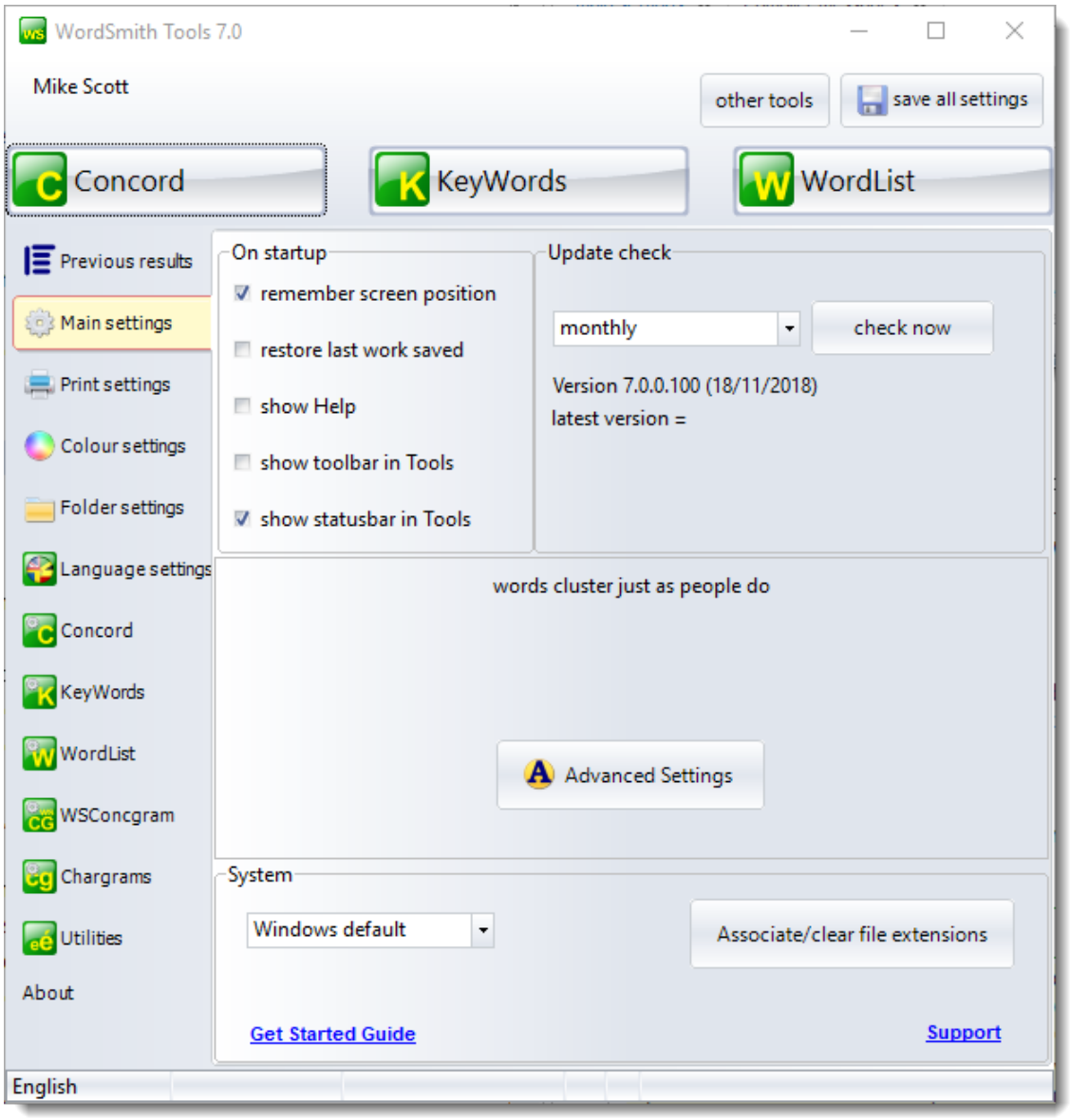


Choose and save 110 settings concerning:

- font 118
- colours 114
- folders 539
- tags 157
- general settings 113
- match-lists 82
- stop lists 141
- lemma lists 322
- text and language settings 145
- Concord Settings 253
- KeyWords settings 297
- WordList settings 367
- advanced user specific settings 120
- index file settings 330

5.19.4 main settings

Found in the WordSmith Tools Controller 5.



Startup

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

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

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

Check for updates

WordSmith can be set to check for updated versions weekly, monthly or not at all. You may freely update your version within the version purchased (e.g. 7.0 allows you to update any 7.x version until 8.0 is issued).

Toolbar & Status bar

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

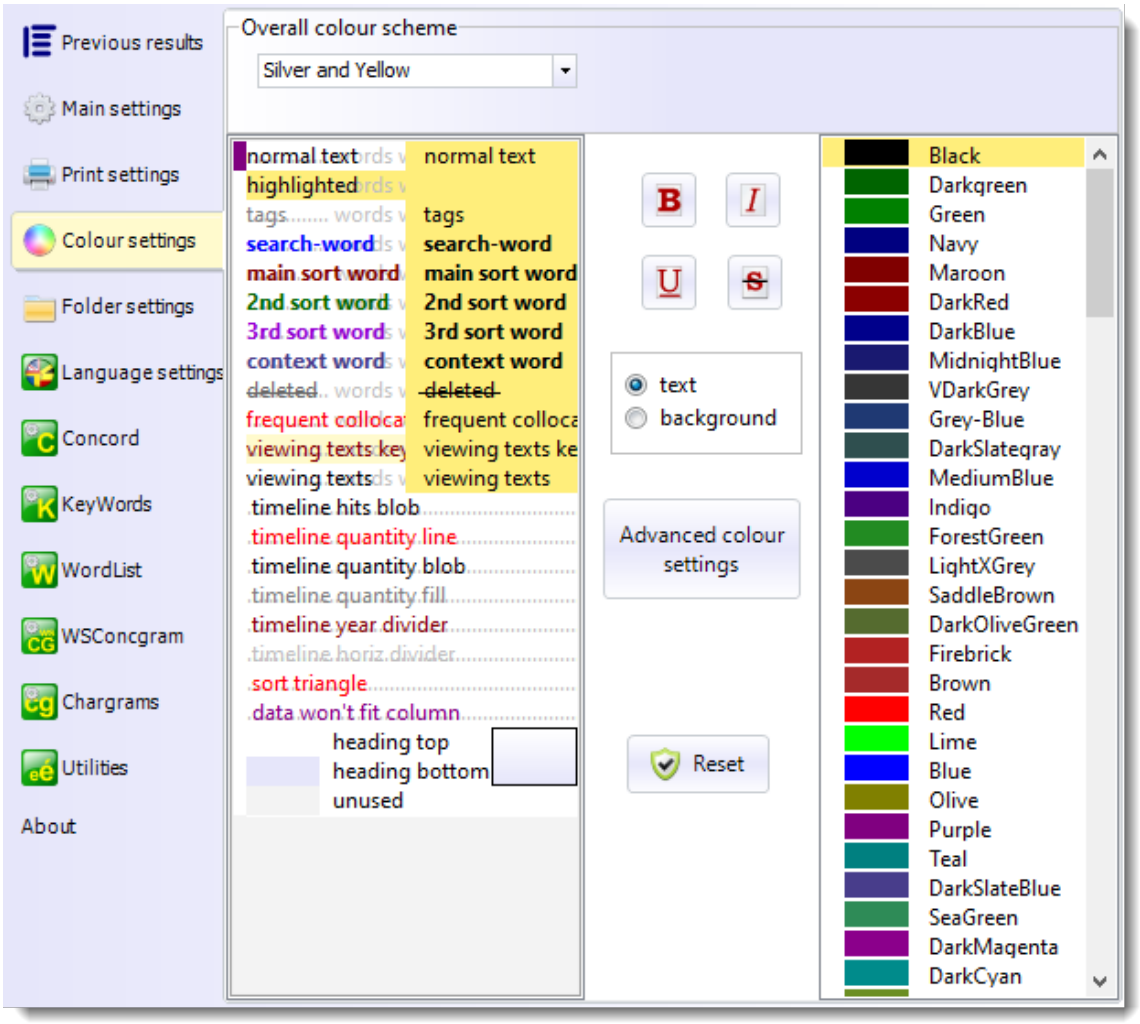
System

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

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

5.19.5 colours

Found in main Settings menu in all Tools and Main Settings in the Controller 51. Enables you to choose your default colours for all the Tools.



Overall colour scheme

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

If you have a word list, concordance or key word list open you can see the effect of any changes as you try them out.

Specifics

To alter colours, first click on the wording you wish to change, then click on a colour in the colour box. The *text* and *background* radio buttons determine whether you're changing foreground or background colours. You can press the Reset button if you want to revert to standard defaults.

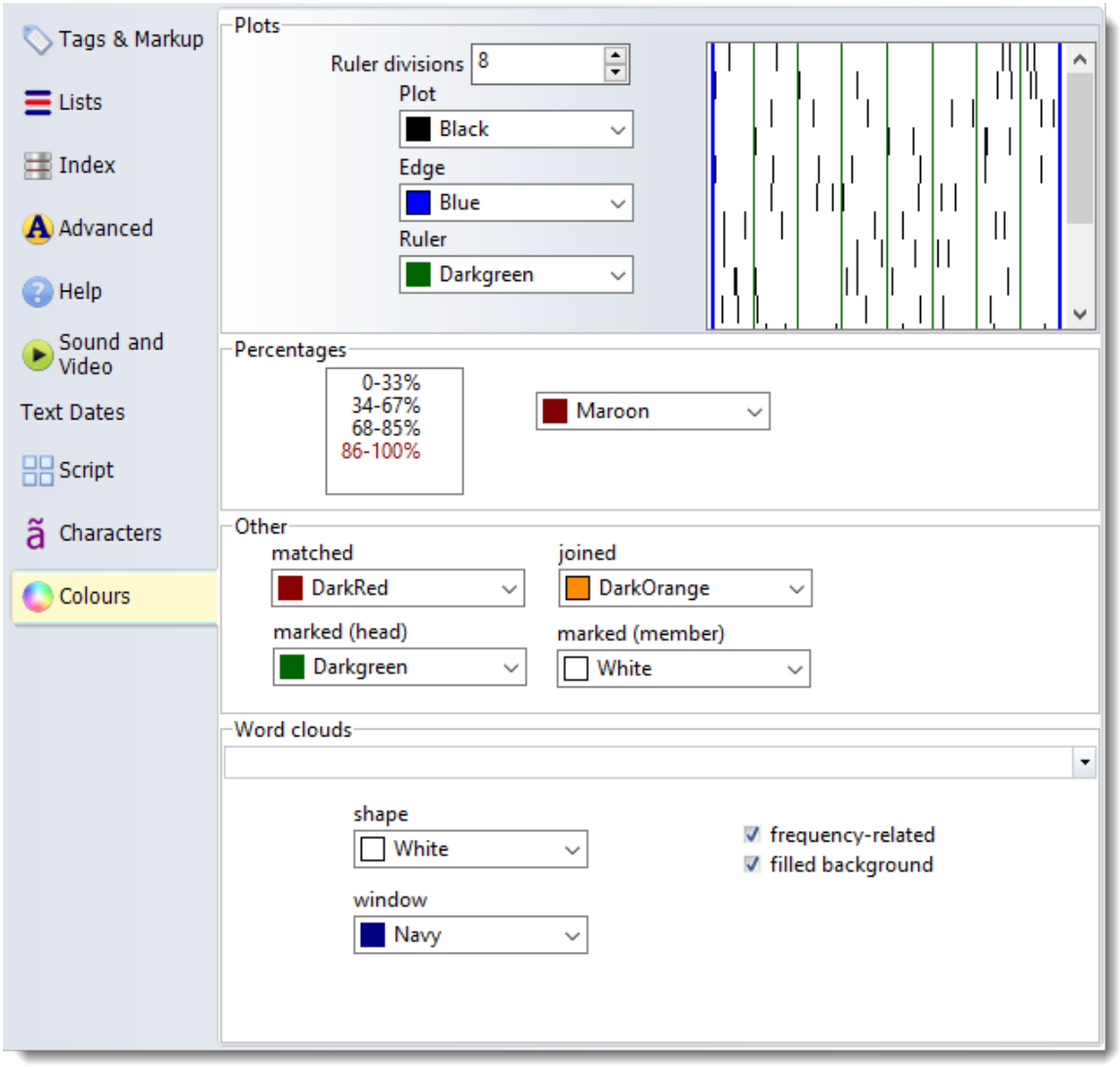
Press the buttons to force your search words to be shown in **bold**, *italics*, etc.

The last few sections concern the overall look of WordSmith's grids, such as the colour of small triangles showing ascending or descending sort, gradient shading in the headings and numbers.

The same colours, or equivalent shades of grey, will appear in printouts, or you can set the printer 113 to black and white, in which case any column not using "plain text" colour will appear in italics (or bold or underlined if you have already set the column to italics).

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

The Advanced colour settings button lets you set some other colours for word clouds, plots, percentage columns, lemmatised, matched and marked words:

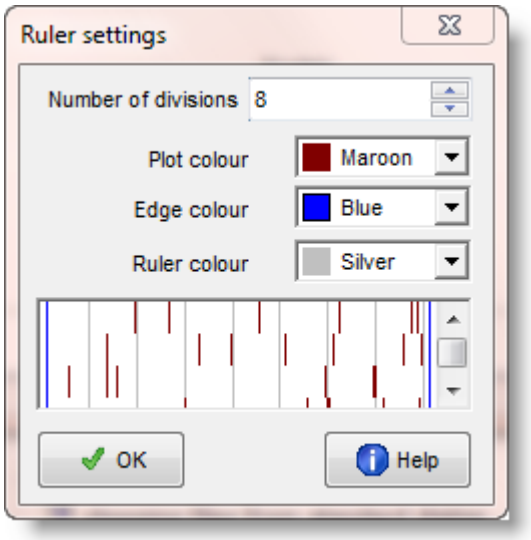


See also: Advanced colour settings 117, Column Layout 60 for changing the individual colours of each column of data, Colour Categories 35, Word Clouds 152.

5.19.5.1 advanced colour settings

Ruler

Advanced colour settings lets you set colours and plot divisions for the ruler:

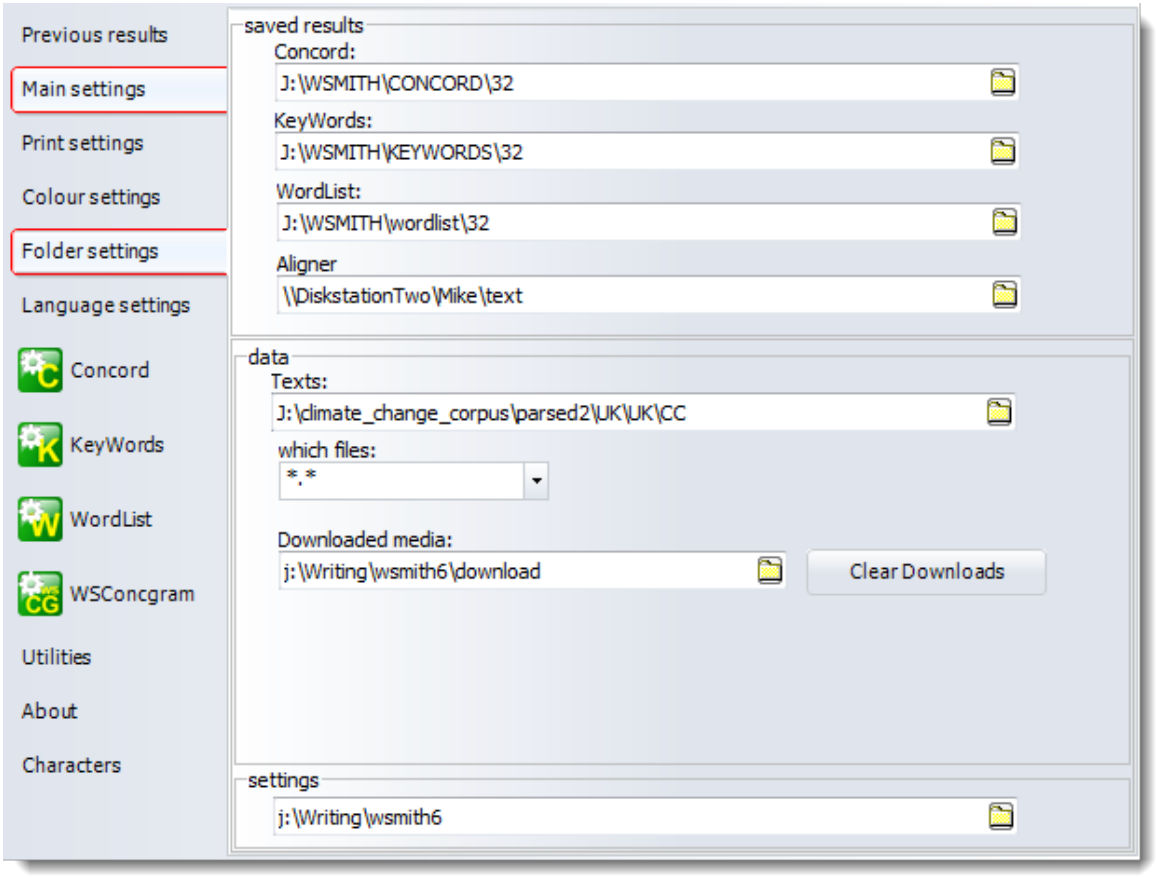


Word Clouds

These settings allow to to choose how each word will be displayed, e.g. within rectangles or circles. The colours of the words and the word cloud window are set in the List colours section above. If you choose *frequency-related colours* the colours vary according to each word's frequency. *Filled background* shows a coloured background if you have no shape (*just the word*) selected.

5.19.6 folder settings

These are found in the main Controller.

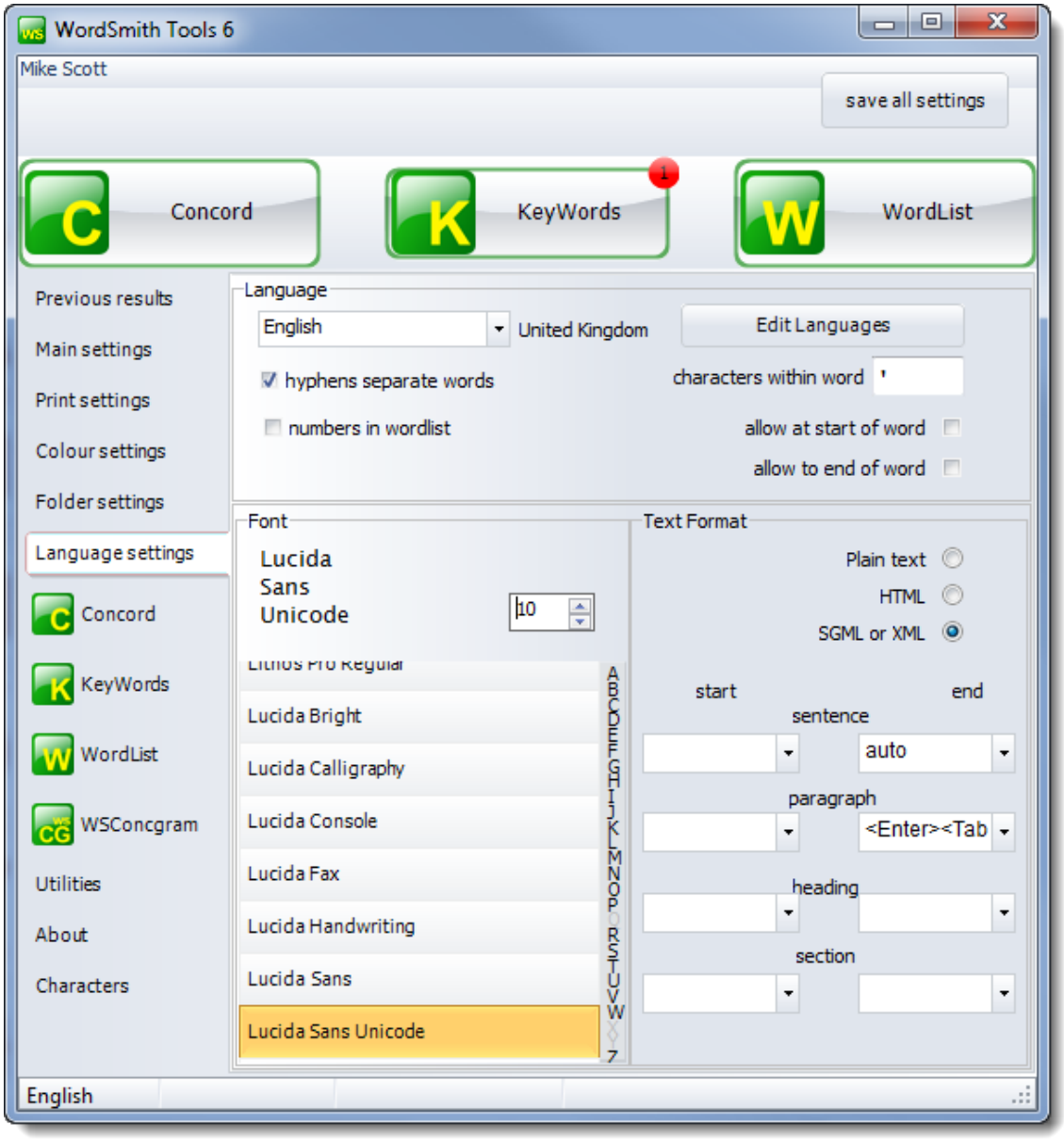


The settings folder will be default be a sub-folder of your My Documents folder but it can be set elsewhere if preferred.

5.19.7 fonts

Found by choosing *Settings* | *Font* in all Tools or via *Language Settings* in the Controller 5. Enables you to choose a preferred Windows font and point size for the display windows and printing 88 in all the WordSmith Tools suite. Note that each language 77 can have its own different default font.

You need to have a) the source texts in a suitable format 527 for WordSmith and b) a font which can display all the characters in your texts.



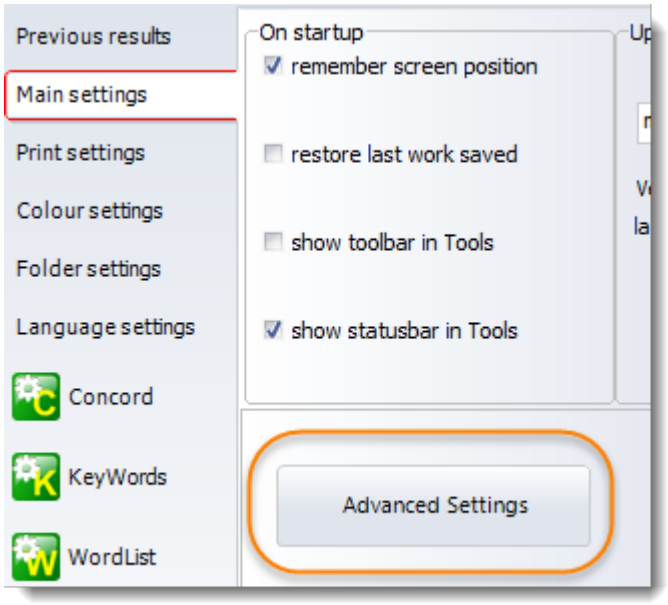
If you have data visible in any Tool, the font will automatically change; if you don't want any specific windows of data to change, because you want different font sizes or different character sets in different windows, minimise these first.

To set a column of data to bold, italics, underline etc., use the layout 60 option .

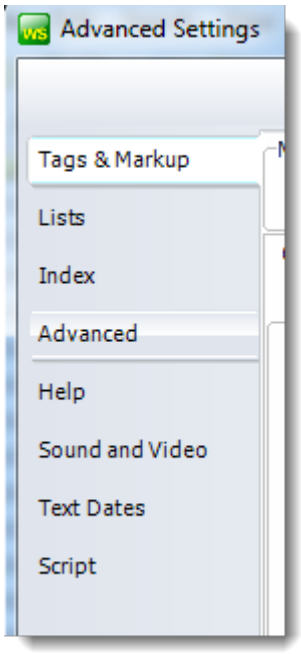
WordSmith Tools will offer fonts to suit the language 77 chosen in the top left box. Each language may require a special set of fonts. Language choice settings once saved can be seen (and altered, with care) in **Documents\wsmith7\language_choices.ini**.

5.19.8 advanced settings

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



and open up a whole new set of options

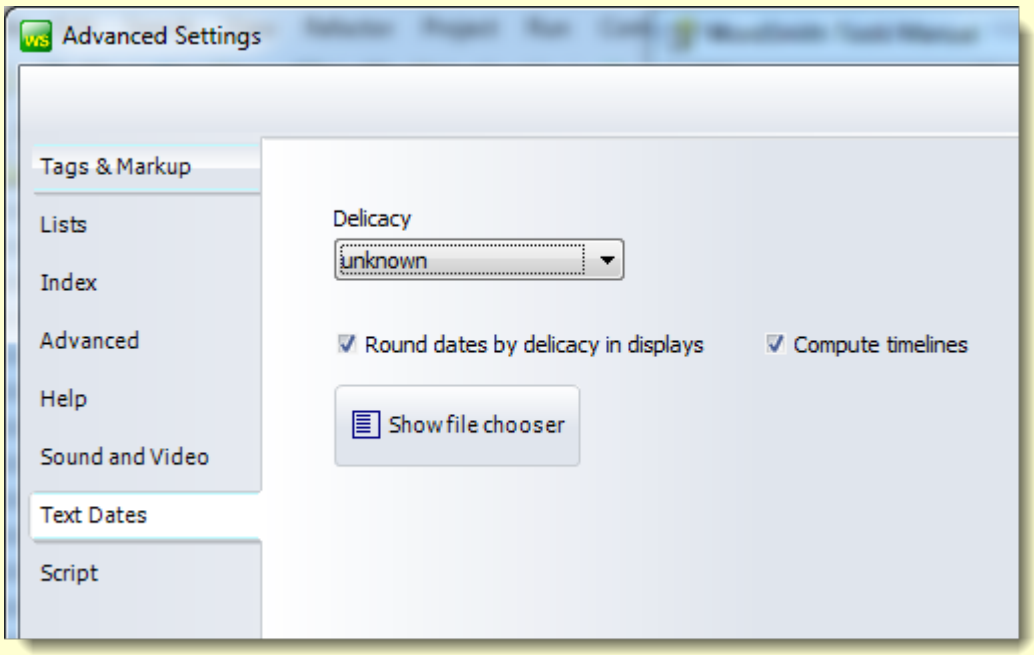


Tags & Markup 157
Lists 362
Index 366

Help
Sound and Video
Text Dates
Scripts 123
Characters
Colours

Text Dates

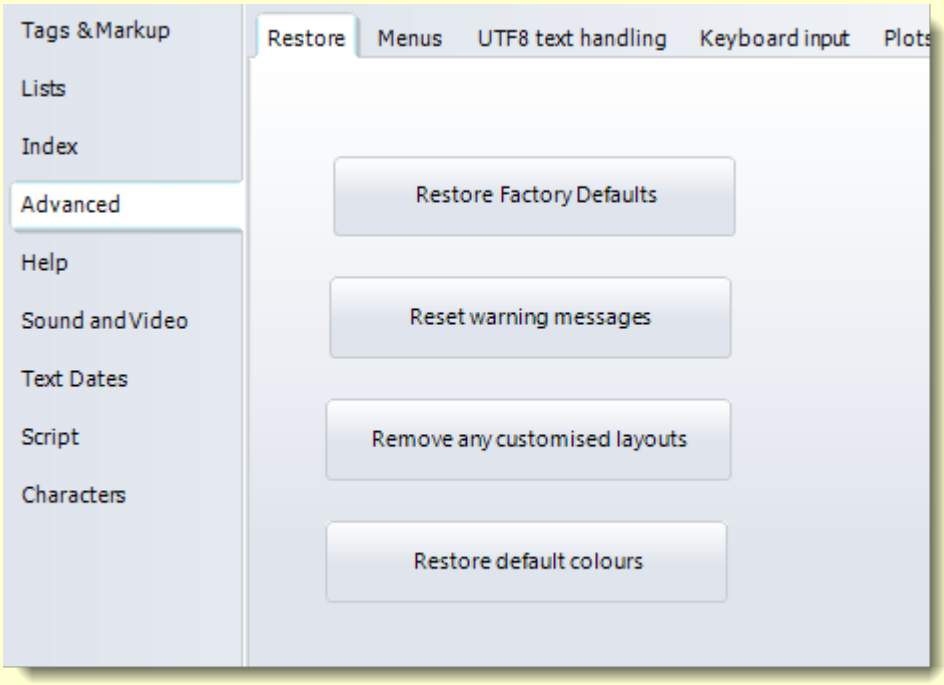
Text dates can be set to varying levels of delicacy, depending on the range of text file dates chosen.



See also: using text dates 149

Advanced section (menus, clipboard, deadkeys etc.)

Restore



These buttons let your restore WordSmith to a previous state.

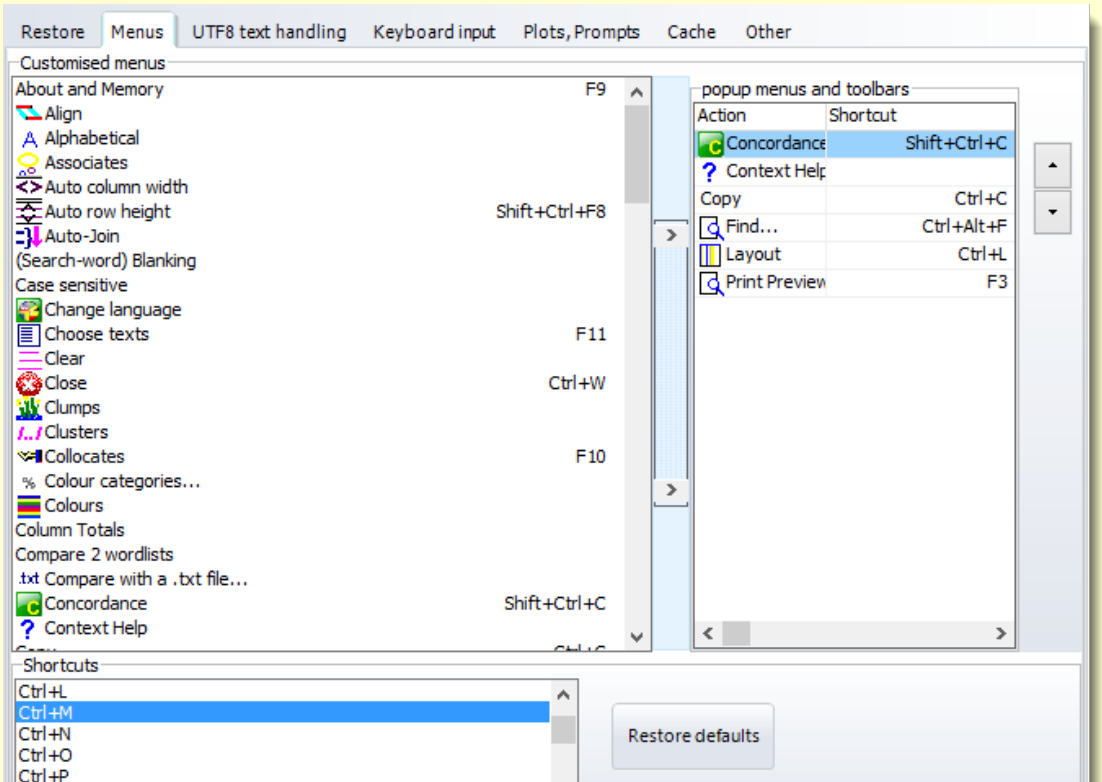
Here you may press a button to restore all factory defaults, useful if your settings are giving trouble.

Customising menus

You can re-assign new shortcuts (such as Alt+F3, Ctrl+O) to the menu items 548 which are used in the various Tools.

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

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

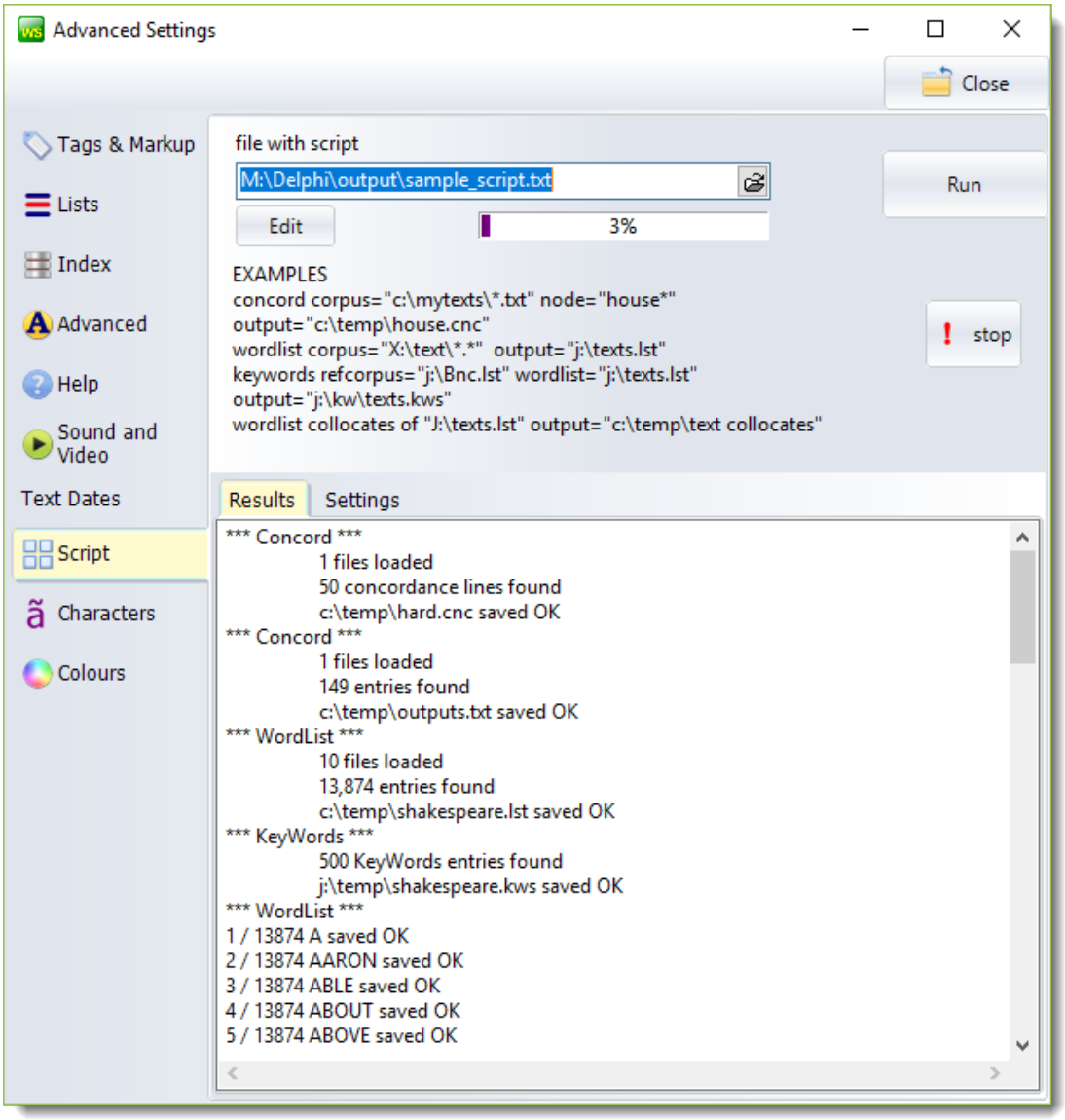


See also : menu and button options 548.

5.19.8.1 scripting

Scripts

This option allows you to run a pre-prepared script. In the case below, **`sample_script.txt`** has requested two concordance operations, a word list, and a keywords analysis. It's done 3% of a further WordList job. The whole process happened without any intervention from the user, using the defaults in operation.



The syntax is as suggested in the EXAMPLES visible above. (There is a **sample_script.txt** file in your Documents\wsmith7 folder). First the tool required, then the necessary parameters, each surrounded by double quotes, in any order. Not case sensitive. Each command must be on one line.

examples

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

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

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

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

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

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

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

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

wordlist-index corpus="x:\text\shakes\oll\txt\tragedies*.txt" clusters = "3" output="c:\temp\shakespeare_tragedies_3s.lst" excluded_subfolders="*_characters" include_subfolders="true"

computed an index of shakespeare tragedies, looked at sub-folders but excluded any containing **_characters**, then computed clusters 3 words in length and saved the word list as **shakespeare_tragedies_3s.lst,**

other output formats (e.g. Excel .xlsx)

TXT_format="true"
Excel_format="true"
RTF_format="true"
XML_format="true"

Saves the output instead, as a **txt, .xlsx, .rtf**, or **.xml** file for use with other software. **RTF_format** is slower than the others.

other options: 1_at_a_time, fetch, show

1_at_a_time="true"

fetch="N" (N = a number)
show="N"

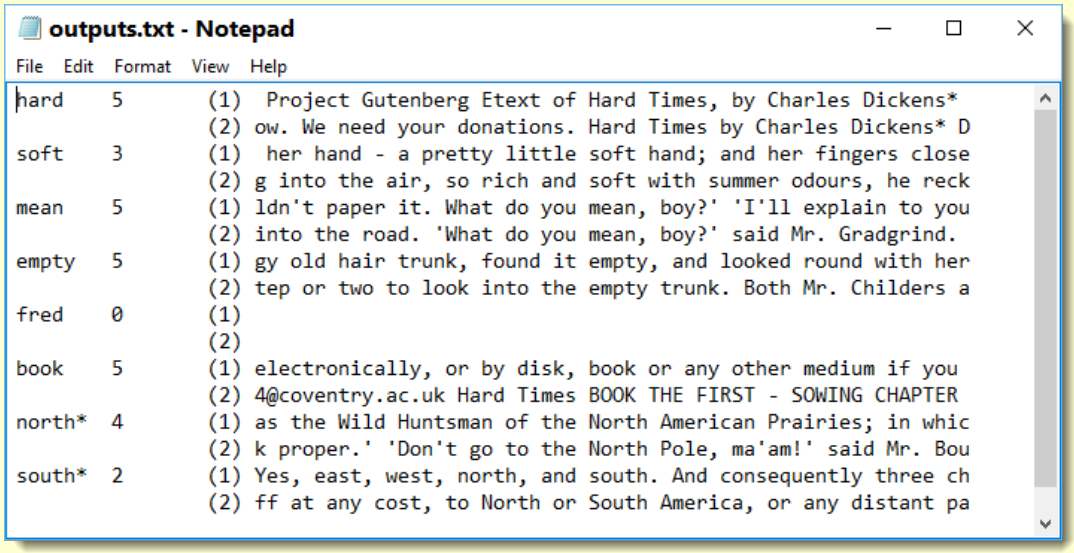
If **1_at_a_time** is true, a word-list will export separate results text file by text file. The **output** needs to be a folder, not a file-name.

If **1_at_a_time** is true, Concord will read search words from a text file and save summary results:

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

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

concord corpus="x:\text\dickens\hard_times.txt" node="c:\temp\sws.txt" output="c:\temp\outputs.txt" 1_at_a_time="true" fetch="5" show="2"
using the same text produced this:



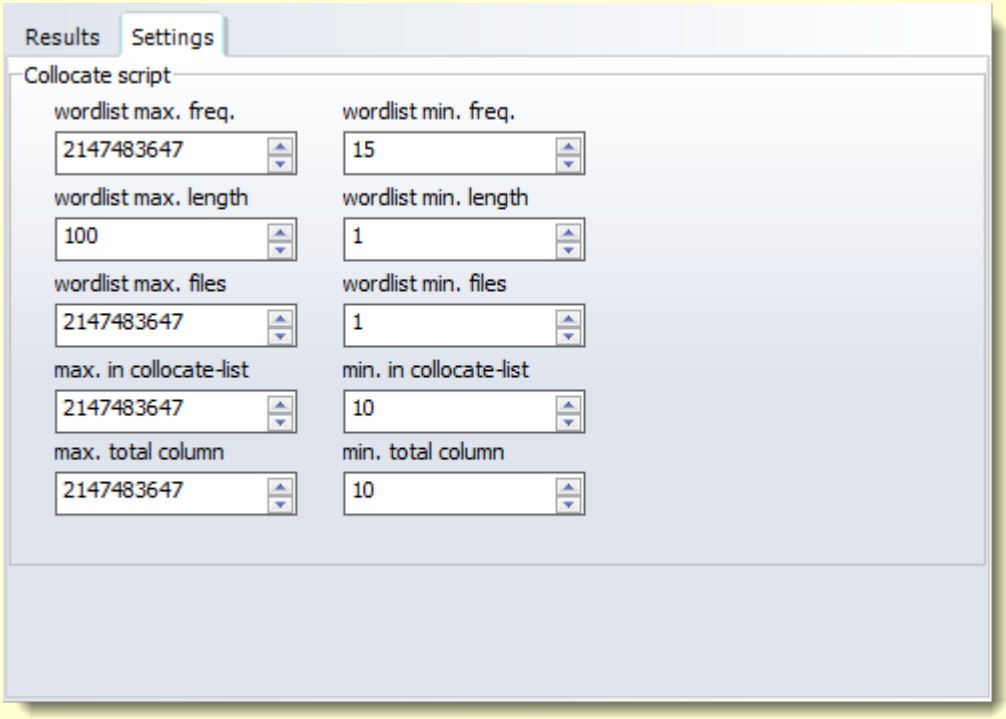
fetch tells WordSmith how many concordance lines to find, **show** tells it how many to show as a sample.

collocate scripts

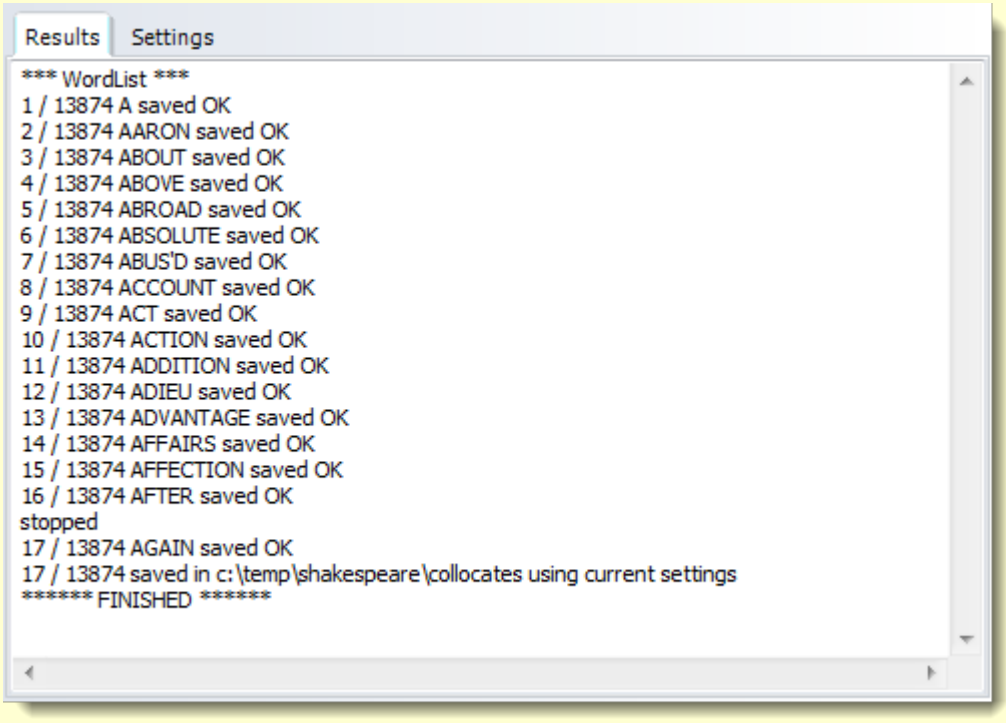
It is also possible to run a script requesting the collocates of each word in a word-list. This syntax

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

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



Results look like this:



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

list of syntax terms

term	use
1_at_a_time	one text-file at a time
clusters	number of words in a cluster
collocates	computes collocates
Concord	use Concord tool
corpus	specifies your text files
Excel_format	save as **.xlsx**
excluded_subfolders	do not include specific sub-folders
fetch	limits the number of concordance lines
include_subfolders	include sub-folders of the main corpus folder in getting text files
KeyWords	use KeyWords tool
node	specifies the search-word
output	specifies where to save results
refcorpus	reference corpus
RTF_output	saves as **.rtf** file
show	limits number of concordance lines to show in output
TXT_format	save output as a **.txt** file
WordList	use WordList tool
WordList-index	make a WordList index
XML_format	output is saved in **.xml** format

See also : drag and drop 534

5.19.8.2 custom processing

This feature -- which, like API 523, is not for those without a tame programmer to help -- is found under *Main Settings* | *Advanced*.

The point of it...

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

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

How to do it...

Choose your **.dll** file (it can have any filename you've chosen for it) and check one or more of the options in the Advanced page. You will need to call standard functions and need to know their

names and formats. It is up to you to write your own .dll program which can do the job you want. This can be written in any programming language (C++, Java, Pascal, etc.).

An example for lemmatising a word in WordList

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

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

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

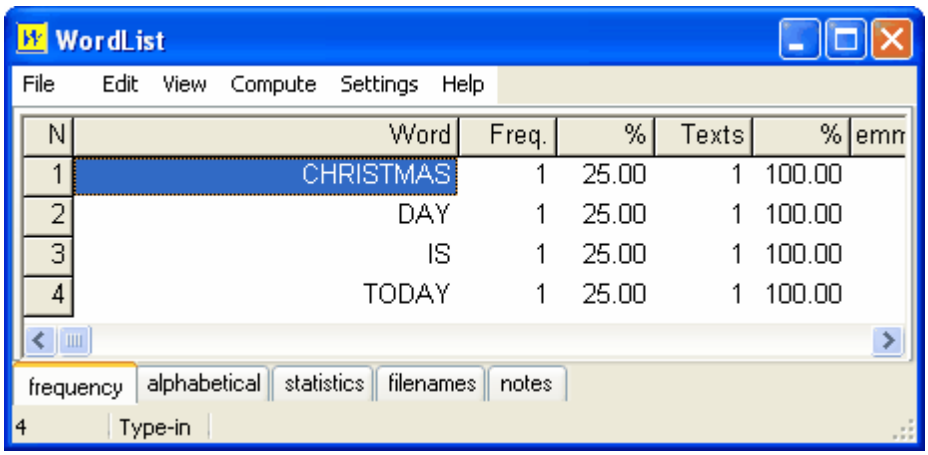
The language_identifier is a number corresponding to the language you're working with. See List of Locale ID (LCID) Values as Assigned by Microsoft .

So the "original" (sent by WordSmith) can be a PANSICHAR (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 PANSICHAR/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.



Source code

The source code for the .dll in Delphi is this

```
library WS7WordSmithCustomDLL;

uses

  Windows, SysUtils;

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

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

 In this example, EASTER gets changed to CHRISTMAS.
}

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

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

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

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

function HandleConcordanceLine
 (source_line : pointer;
```

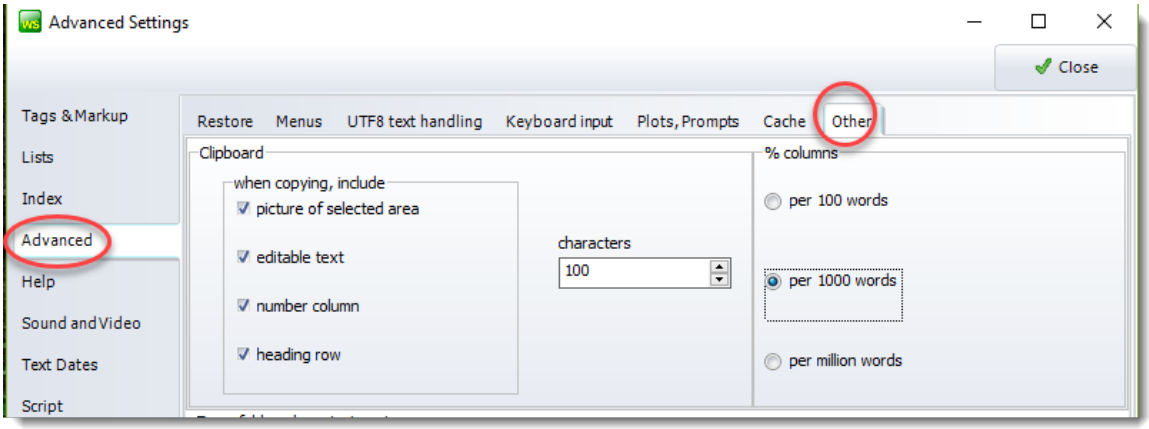
See also : API 523, custom settings 165

5.19.8.3 percentage columns

Traditionally WordSmith has shown columns headed % as percentages. This is a familiar concept but does show tiny numbers. Accordingly in the Advanced Settings you can set the display to

- per 100 words
- per 1000 words
- per million words

Find it in *Advanced Settings* | *Advanced* | *Other*.



This setting will be saved permanently if you save your settings 110.

Here is an example where a KeyWords display has the setting at per 1,000 words:

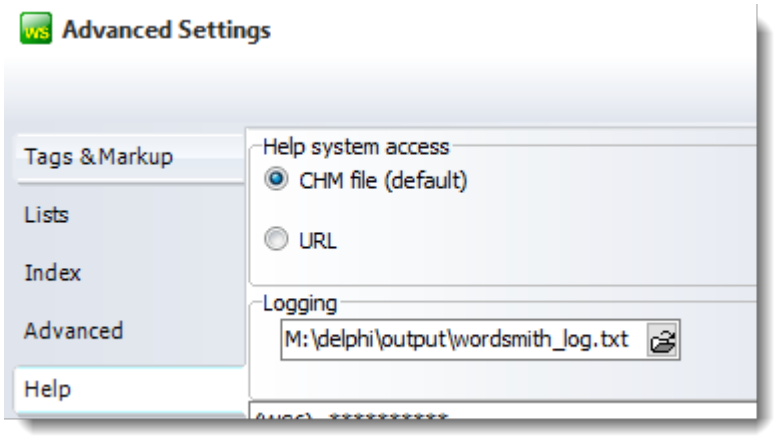
Key word	Freq.	per 1000 words	Texts RC. Freq.		per 1000 words	BIC
ROMEO	115	4.77	1	130	0.15	481.57
TYBALT	47	1.95	1	52	0.06	190.11

See also: layout 60 of the data in rows and columns.

5.19.8.4 Help_and_logging

Help system access

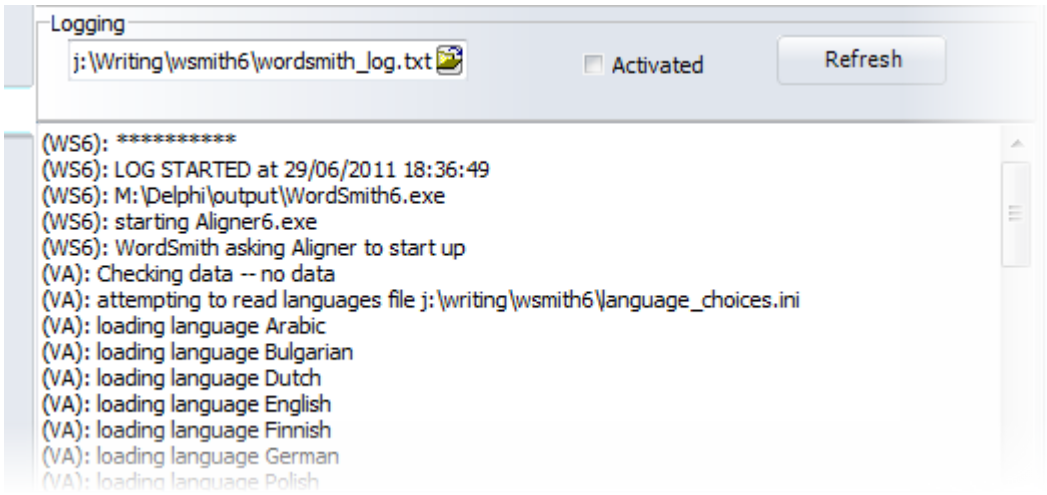
On a network, it is commonly the case that Microsoft protects users to such an extent that the usual **.CHM** help files show only their table of contents but no details. Here you can set the WordSmith help to access the local CHM file or the online Help at the WordSmith URL.



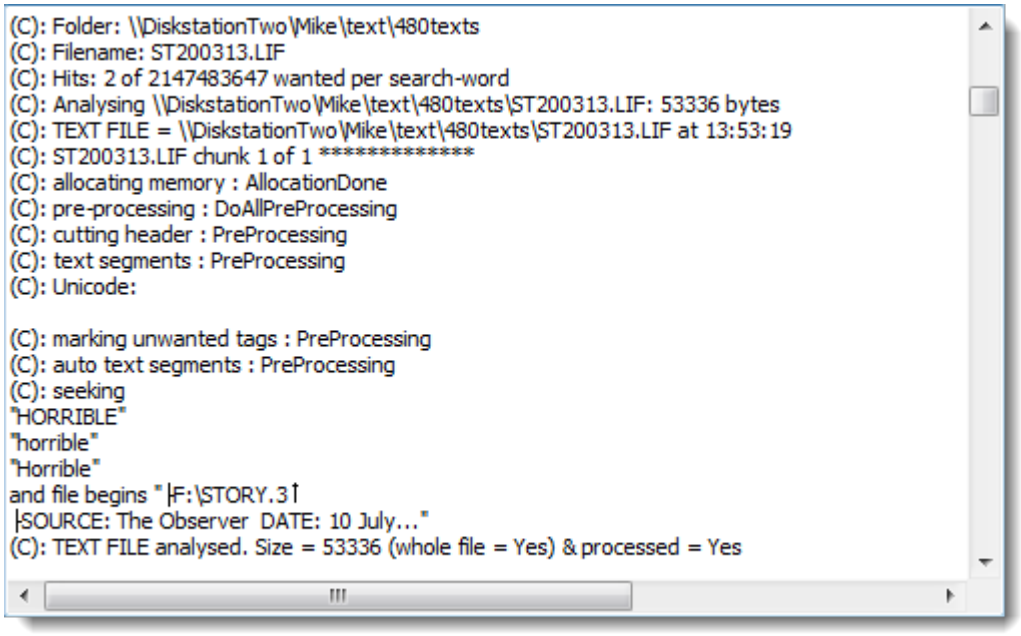
Logging

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

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

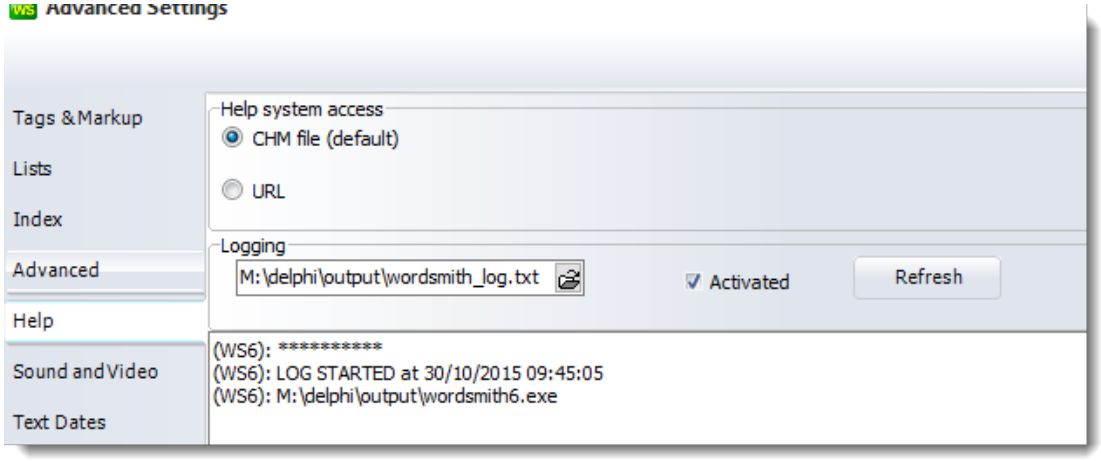


And here in a Concord process we see some details of the text files being read and processed, seeking the search-word **`horrible`**:



The most straightforward way to use logging is

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



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

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

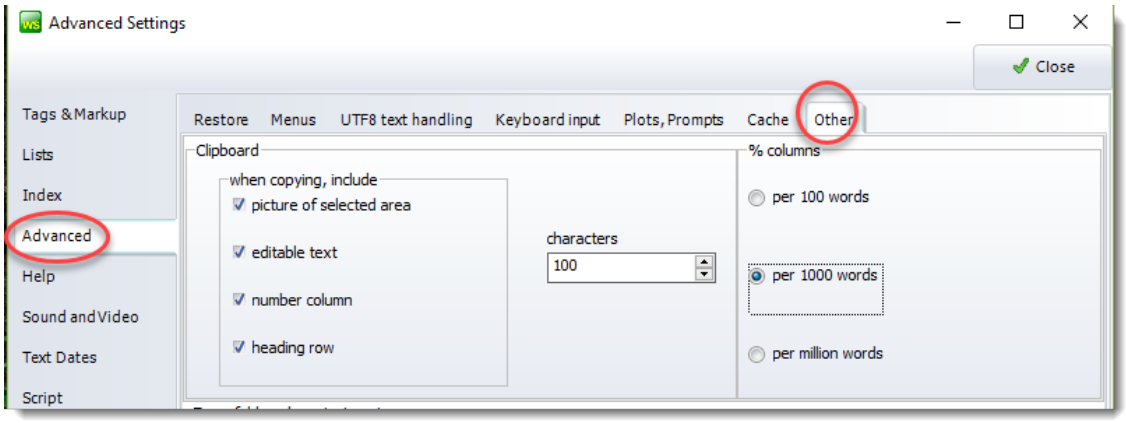
If you want to log as WordSmith starts up, start in from the command line with the parameter **/log**:

Start | Run | Cmd <Enter> **| cd\wsmith7** <Enter> **| wordsmith7 /log** <Enter>
(or **wordsmith7 /log C:\temp\WSLog.txt** to force use of **C:\temp\WSLog.txt**. If you do that, make sure the folder exists first.)

See also: emailed error reports 525.

5.19.8.5 clipboard

Here is where you can change default settings for copying data to the clipboard[529].



The number of characters only applies when copying as editable text. See also: clipboard[529].
Settings changed here can be saved permanently if you Save all settings[110].

5.20 source text(s)

The point of it...

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

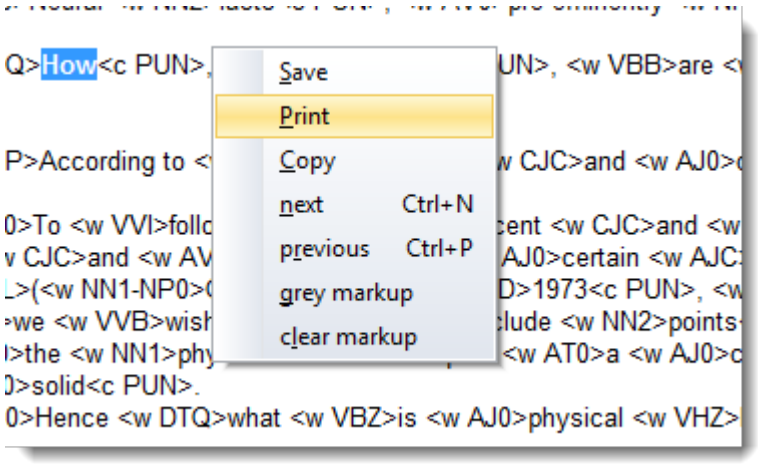
How to do it

The **Concord** and **KeyWords** tools both have areas which can show the source texts which your data was produced from, visible by choosing the *source texts* tab, if your texts are still where they were when the data analysis was done. (If they have been moved you can try editing the *Filenames*[102] data to correct this.)

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

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

Menu options (right-click to see these)



Copy, Print, Save

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

Highlight words

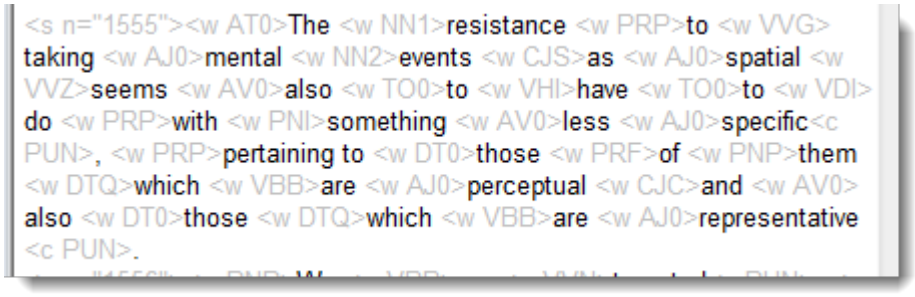
Gets the key words (or search-words in Concord) highlighted so you can see each in its context and can go from one to the next.

Next, Previous

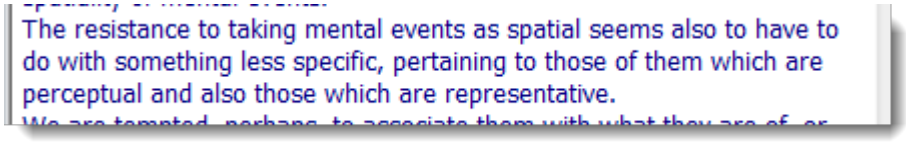
These jump you through the text one highlighted word at a time. You should see how many highlighted there are in the status bar.

Grey markup, Clear markup, Restore markup

Grey markup lets you grey out all < > sections.

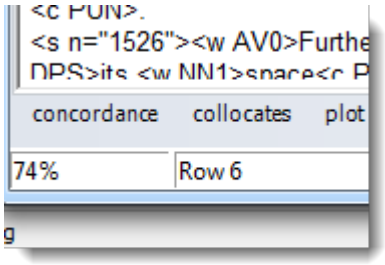


Clear markup simply cuts the tags out.



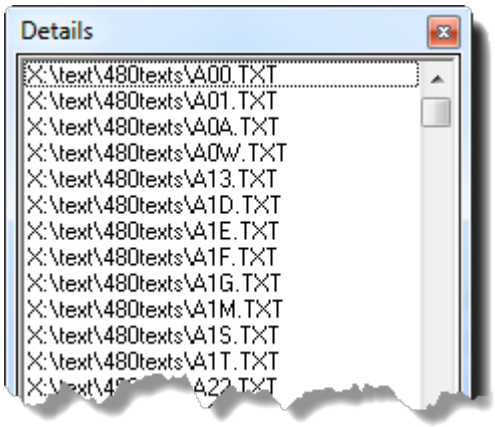
Once you have cut out markup, the *clear markup* option changes to *restore markup* (needed if Concord is to jump to the correct location).

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

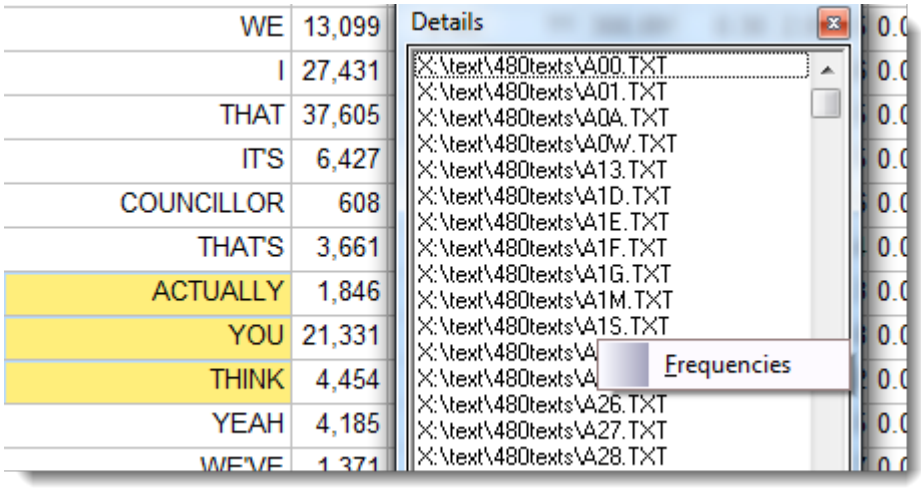


Double-clicking the status bar gives you a chance to stop the process.

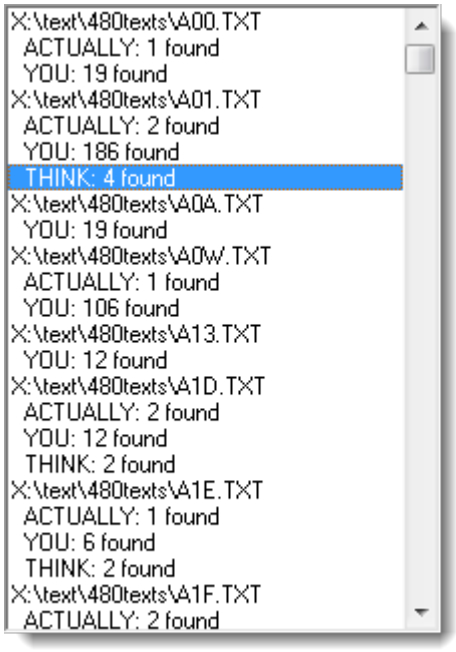
KeyWords List of Source Texts



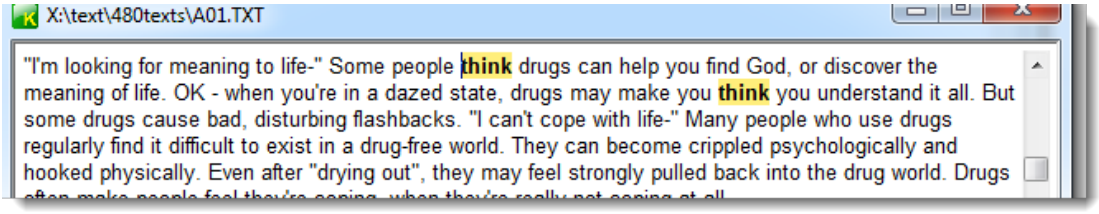
If you right-click this window, you get a chance to see which texts contain which key words



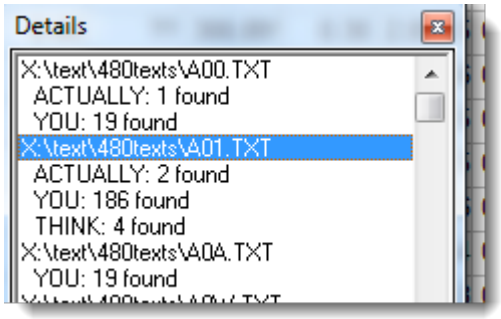
by clicking *Frequencies*, giving results like this:



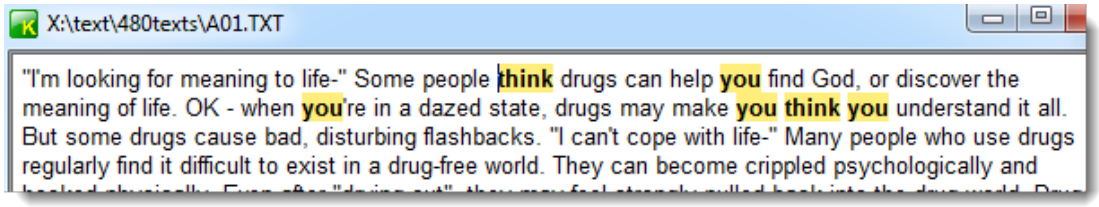
and if you double-click a highlighted word (**THINK** in the example above), you will be shown the source text (**A01.txt**) with that word highlighted.



If you simply click the file-name



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

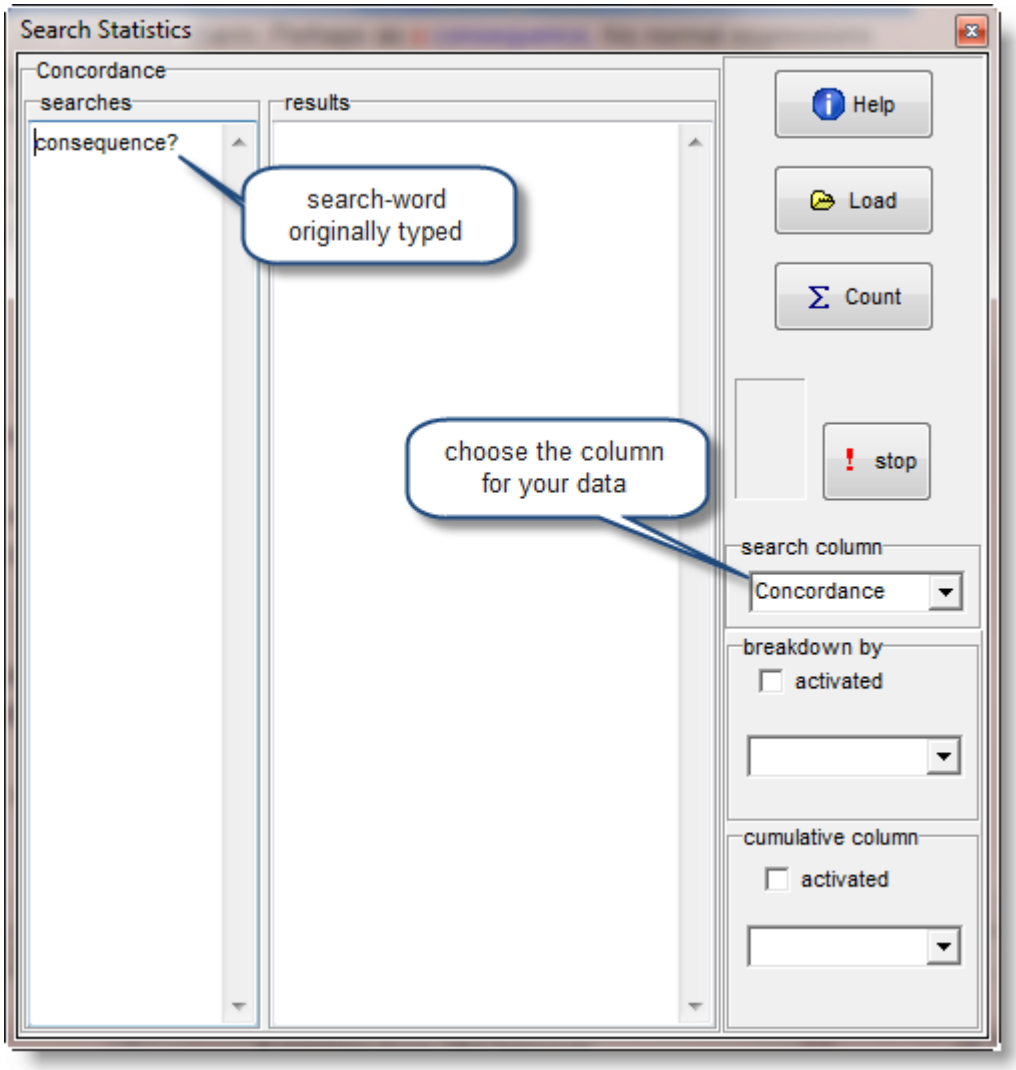


See also: modify source texts 182

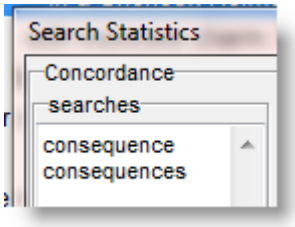
5.21 breakdowns

The idea is to be able to break down your data. For example, you've just done a concordance of **consequence?** which has given you lots of singulars and lots of plurals and you want to know how many there are of each.

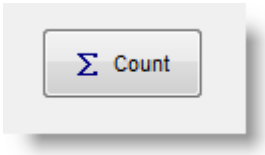
Choose *Breakdown* in the *Compute* menu.



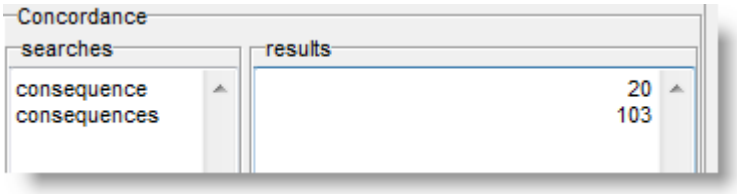
The *searches* window will at first contain a copy of what you typed in when you created the concordance. To distinguish between singular and plural, change that to



and press Count;



assuming that *search column* has *Concordance* selected, you will get something like this:



Load

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

Search Column

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

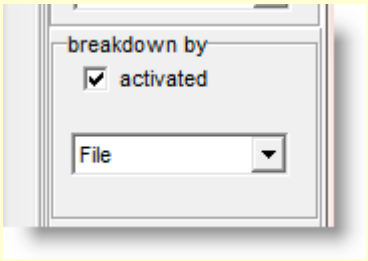
Cumulative Column

This adds up values from another column of data. See the example from WordList 359.

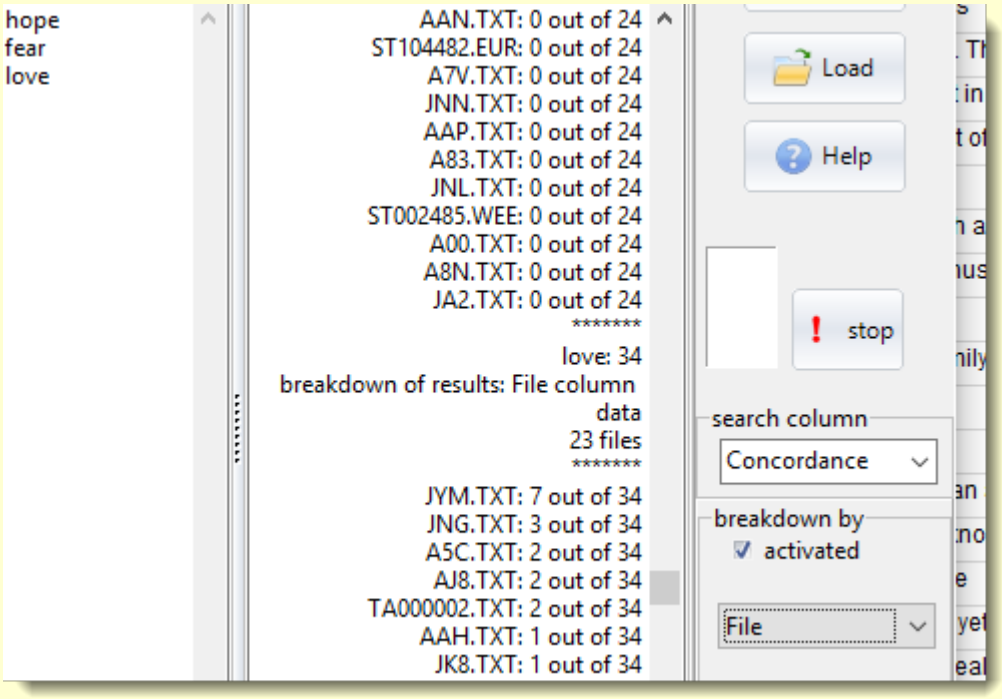
Further breakdown of your results

The idea here is to be able to break down your results further, using another category in your existing concordance data, such as the files the data came from. In our example, we might want to know for consequence and consequences, how many of the text files contained each of the two forms in the concordance lines gathered.

To generate the breakdown, activate it and choose the category you need.



The results window could now show something like this



There's a long list of text file results for each search-word typed in, *hope, fear, love*. The picture shows the last few entries for *fear* and the first few for *love*. That word was found 7 times in the concordance lines from text file **JYM.TXT**. (Note: the concordance was of *hope*, not *love*. There might be more than 7 tokens of love in that text file, the search was on the concordance lines gathered.)

See also: Concord breakdowns 200

5.21.1 count data frequencies

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

See also: distinguishing consequence from consequences 138, save text file data 122, frequencies of suffixes in a word list 359, compute new column of data 65.

5.22 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*.

Prepare and load up your stop list first, then make your word lists using it. (To use a list to remove items from an existing word list, see the Match list 82 function.)

To use stop lists, you first prepare a file, using **Notepad** or any plain text word processor, which specifies all the words you wish to ignore. Separate each word using commas, or else place each one on a new line. You can use capital letters or lower-case as you prefer. You can use a semi-colon for comment lines. There is no limit to the number of words. Stop lists do not use wildcards (match-lists 82 may).

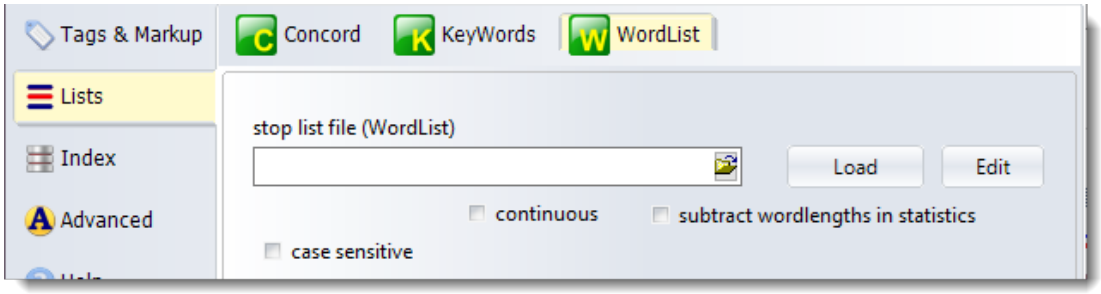
There is a file called **stop_wl.stp** (in your **\wsmith7** folder) which you could use as a basis and save under a new name. You'll also find **basic_English_stoplist.stp** there, based on top frequency items in the BNC. Or just make your own in Notepad and save it with **.stp** as the file-extension. If that is difficult, rename the **.txt** as **.stp**.

Example

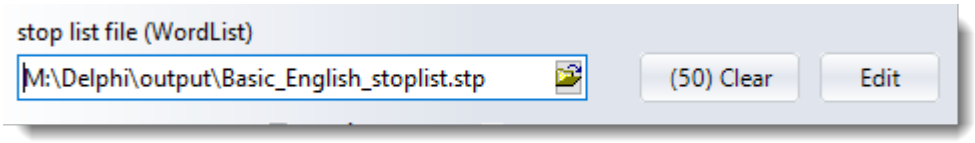
```
; My stop list for test purposes.
THE,THIS,IS
IT
WILL
<VVN>SEEN
```

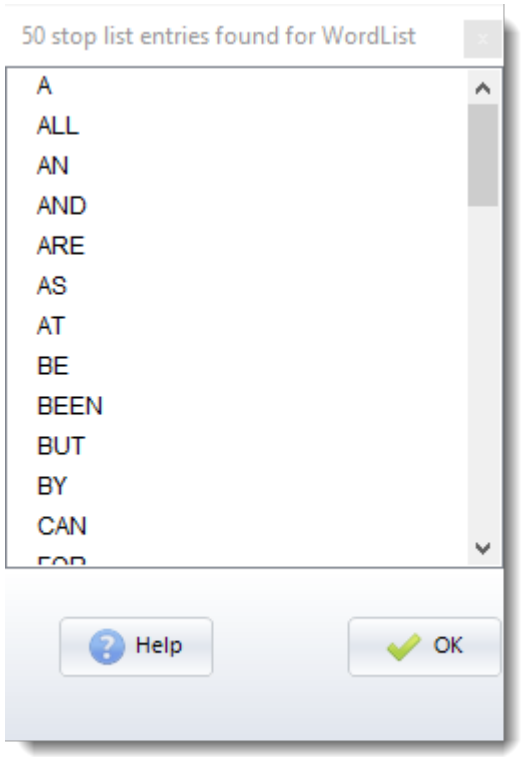
You may put angle-brackets in a stop list. So **<VVN>SEEN** (that is **seen** as a past participle) will remove any cases of precisely that sequence if it is in your texts. It will only stop **seen** if marked with **<VVN>** in the corpus.

Then select *Stop List* in the menu to specify the stop list(s) you wish to use. Press Load to read it in. Separate stop lists can be used for the **WordList, Concord** and **KeyWords** programs.If you wish always to use the same stop list(s) you can specify them in **wordsmith7.ini** as defaults 110.



To choose your stop list, click the small yellow button in the screenshot, find the stop list file, then press *Load*. You will see how many entries were correctly found and be shown the first few of them. Here the Load button now shows a number, so it is in effect: that is, the 50 words in the list will be stopped from being included in a word list. Press Clear to deactivate it.





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

Case Sensitive

If you have made a case-sensitive word list 371, you will probably want the stop list to be case sensitive too.

continuous

Normally, every word is read in while making the word list and stored in the computer's memory without checking whether it's the stop list. Eventually the set of words is checked in your stop list and omitted if it is present. That is much quicker. However, it means that for the most part, any statistics 350 are computed on the whole text, disregarding your stop list.

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

subtract wordlengths in statistics

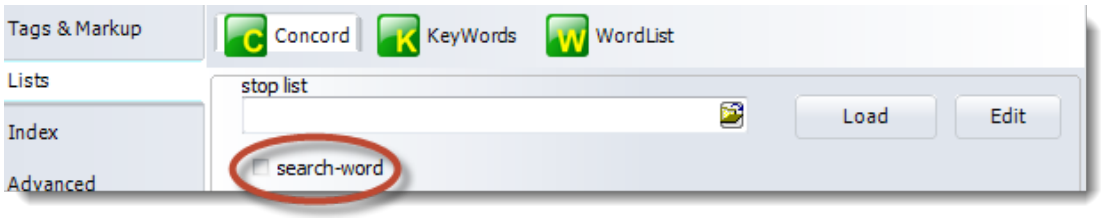
If you have not chosen continuous processing as explained above, you may want the statistics of your word list to attempt to deal in part with the stop list work done. With this choice, after the word list is computed, all the statistics concerning the number of types and tokens and 3-letter, 4-letter words etc. will be adjusted for the overall column (but not for the column for each single text) in your statistics 350.

See Match List 82 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 484.

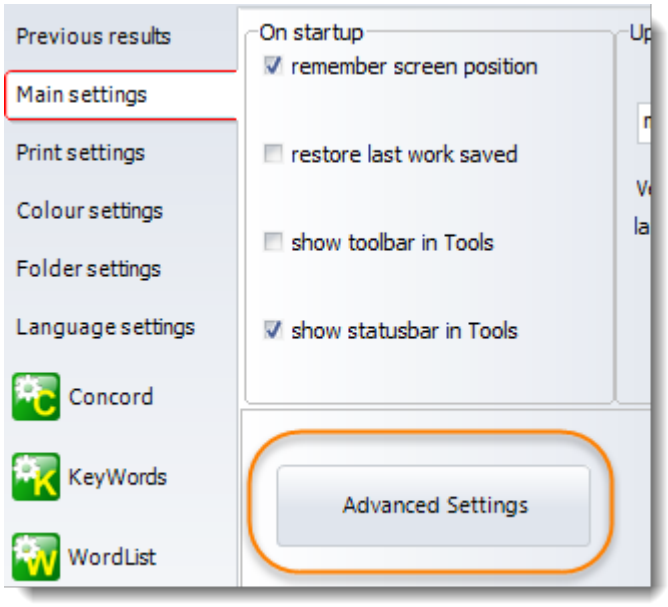
stop lists in Concord

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



Stop lists

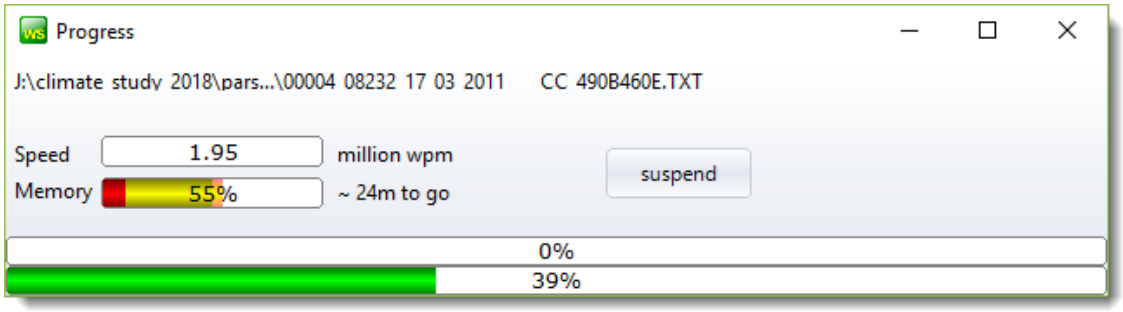
... are accessed via an Advanced Settings button in the Controller



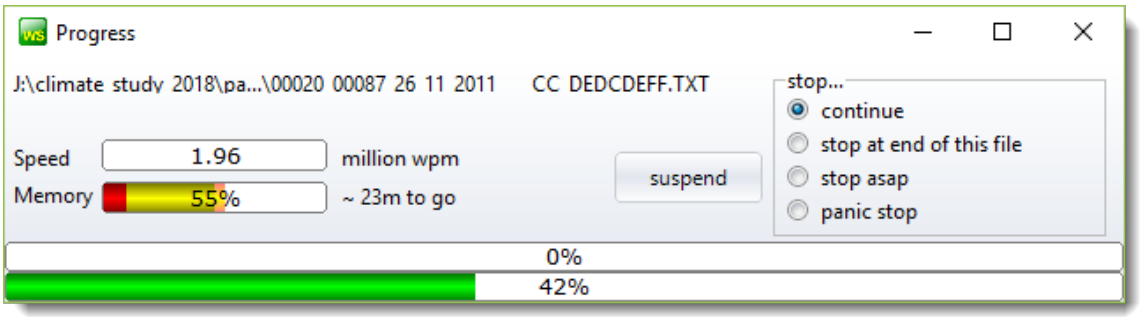
See also: Making a Tag File 172, Match List 82, Lemmatisation 327.

5.23 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.



Pressing *Suspend* offers 4 choices:



continue

... as if you had not interrupted anything

stop at end of this file

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

stop as soon as possible

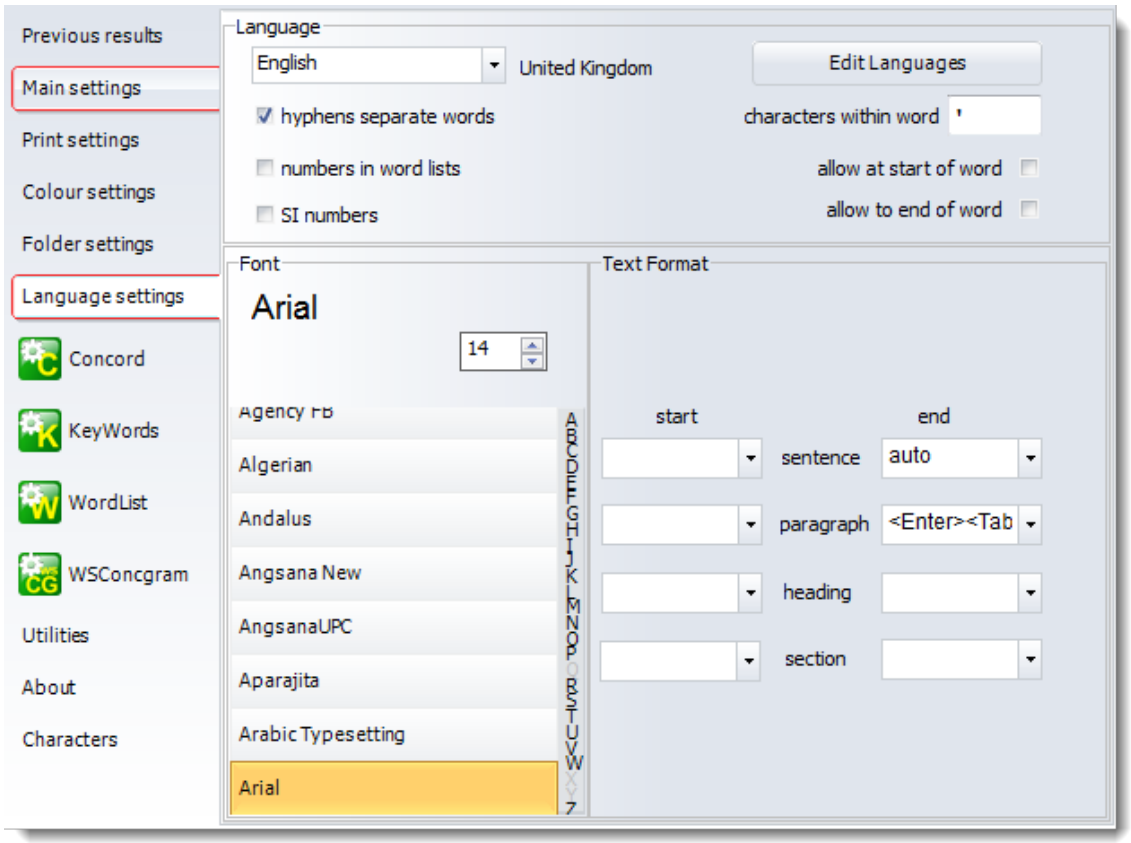
...useful if you're ploughing through massive CD-ROM files. WordSmith will stop processing the current file in the middle, but will retain any data it has got so far.

panic stop

The whole Tool (Concord or WordList, or whatever) will close down and some system resources memory 555 may be wasted. The Controller 5 will not be closed down.

Press *Suspend* again to effect your choice.

5.24 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 9 button.

The choices below may differ for each language:

hyphens and numbers

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

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

characters within word

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

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

Examples:

'	(only apostrophes allowed in the middle of a word)
'%	(both apostrophes and percent symbols allowed in the middle of a word)
'_	(both apostrophes and underscore characters allowed in the middle of a word)

You can include up to 10.

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

(start & end of) headings

For the Tools to count headings, they need to know how to recognise the start and end of one. If your text is tagged 157 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 543 text which is not consistent, using sometimes **</h1>** and sometimes **</H1>**, then use Text Converter 11 to make your texts consistent).

(start & end of) sections

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

(start & end of) **sentences**

If this space contains the word **auto**, the Tools will treat sentences as defined 533 (ending with a full stop, question mark or exclamation mark, and followed by a capital letter), but if your text is tagged 157 e.g. with **<s>** and **</s>**, type those in here. Again, whatever you type is case sensitive.

(start & end of) **paragraphs**

For the Tools to recognise paragraphs, they need to know what constitutes a paragraph start and/or end, e.g. a sequence of two <Enter>s (where the original author pressed Enter twice) or an <Enter> followed by a <Tab>. For that you would type **<Enter><Tab>**. If your text is tagged 157 e.g. with ** ** and ** **, 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 ** **, but you can leave ** ** here as WordSmith will use **</u>** instead if the text has no ** ** in it.

See also: Tagged Text 157, Stop Lists 141, Choosing a new language 9. Processing text in Chinese etc.

5.24.1 characters and letters

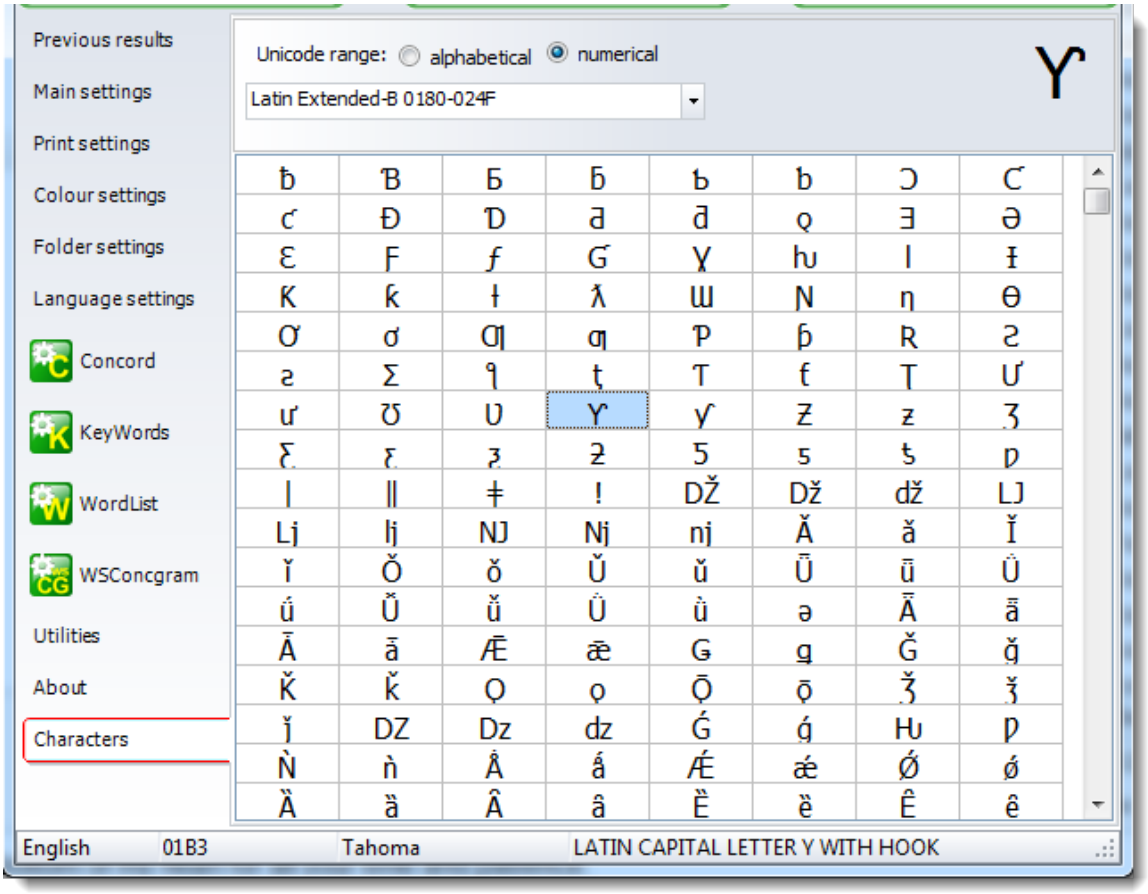
WordSmith uses standard Microsoft routines to distinguish between letters, numbers, punctuation etc. Accented characters 147 are automatically handled. WordSmith works with numerous languages whatever the character set as long as you have a suitable font. Texts in Unicode format allow for a huge variety of characters to be processed in a great number of languages.

You can input characters whatever the keyboard 528.

Se also: wildcards 148

5.24.1.1 accents and other characters

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



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

See also: Copying a character into Concord 528

5.24.1.2 wildcards

Many WordSmith functions allow you a choice of wildcards:

symbol	meaning	examples
*	disregard the end of the word,	tele*
	disregard a whole word	*ness
		happi
		book * hotel
?	any single character (including punctuation) will match here	Engl???
		?50.00
^	a single letter	Engl^^^
#	any sequence of numbers, 0 to 9	$#
		£#.00

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

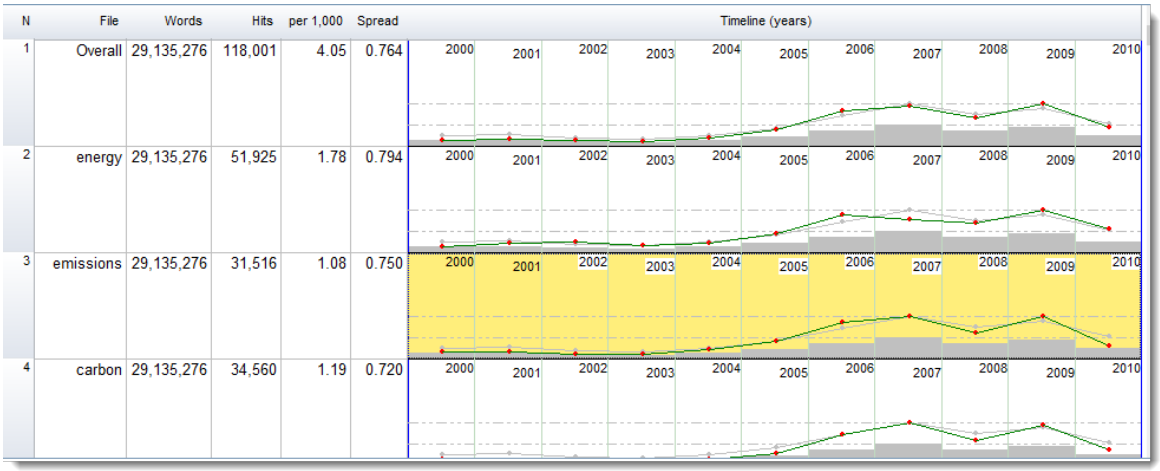
5.25 text dates and time-lines

The point of it ...

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

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

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



Columns

The *Words* column shows the number of tokens, *Hits* shows the number of concordance entries each time-line entry is based on, *per 1,000* is the number of hits per 1,000 tokens and *Spread* is the dispersion value 554.

Time-line display

The first line shows overall data where all results on three search-terms are merged. Each line can be pulled taller or shorter.

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

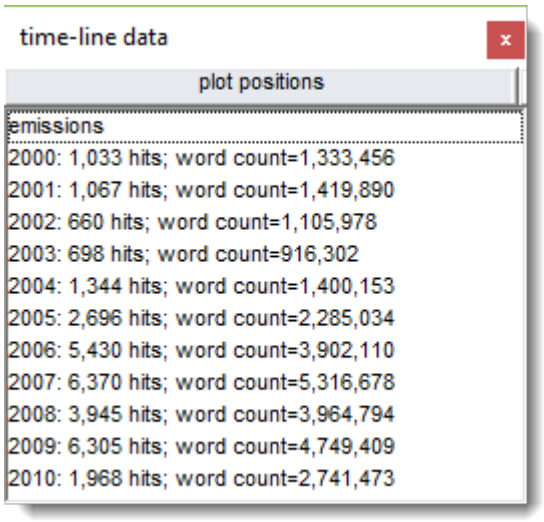
The grey rectangles and the grey graph line both represent the same background information, namely the amount of word-data searched. The difference is merely that the grey line is twice as high as the rectangles below it.

The number of hits in each year is mostly roughly proportional to the amount of text being examined, though in 2006 and 2009 for the term **emissions** it seems that the hit rate was slightly

higher. In the first half of the decade **carbon** was rather under-mentioned in proportion to the amount of climate-change data being studied.

Double-click the time-line entry

To see the detail of an entry, double-click. Here I double-clicked the cell highlighted in the screen-shot above and got this:



See also: choosing text files - setting file dates 56

5.26 view

Show percentages

This menu option (*View* | *Show as %*) lets you toggle between raw numbers and percentages. For example, this collocate listing shows raw numbers,

Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
154	425	13	13	1	4	5	3		399		3	5	4	1
39	63	17	46	2	4	2	5	4		29	7	3	4	3
1	2	1	1		1						1			
2	2	0	2									2		

but if you switch to percentages it shows row percentages.

Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
425	3.1%	3.1%	0.2%	0.9%	1.2%	0.7%		93.9%		0.7%	1.2%	0.9%	0.2%
63	27.0%	73.0%	3.2%	6.3%	3.2%	7.9%	6.3%		46.0%	11.1%	4.8%	6.3%	4.8%
2	50.0%	50.0%		50.0%						50.0%			
2	0.0%	100.0%									100.0%		

in a concordance, the raw numbers for sentence position

4	compared with the warmth of family, **love,** friendship, security and	20
5	In Ariel (15) a miner who has been made redundant heads south to Helsinki and falls in **love.** "	18
6	Many would **love** to know, for instance, whether	3
7	10 single, the old Motown number "**Love** Pains" and, rather curiously,	19

can be seen instead as percentages:

14	compared with the warmth of family, **love,** friendship, security and	83.3%
15	In Ariel (15) a miner who has been made redundant heads south to Helsinki and falls in **love.** "	100.0%
16	Many would **love** to know, for instance, whether	5.9%

5.27 window management

The main WordSmith Tools Controller 5 will be at the top left corner of your screen; it won't take up the full screen size and there's no need to maximize it because it only shows settings. Other Tools probably will be bigger and can usefully be maximized.

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 5 window when using individual Tools.

tile and cascade

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

restore last file

A convenience feature: the last file you saved or retrieved will by default be restored when you re-enter WordSmith Tools. I've kept it to one only to avoid screen clutter! This feature can be turned off temporarily via a settings option or permanently in **wordsmith7.ini** (in your **Documents\wsmith7** folder). You can also generally access your last saved result in any Tool by

right-clicking and choosing last file:

5.27.1 filenames tab

This tab shows the text file name(s) from which your current data comes. You can edit these names if necessary (e.g. if the text files have been moved or renamed 122.) To do so, choose *Edit* | *Searching* | *Search & Replace* ( ).

Afterwards, if you save the results 91, the information will be permanently recorded.

See also: finding source files 537.

5.28 word clouds

The point of it...

Many of the lists in WordSmith offer a word cloud feature similar to those you have probably seen on the web. The idea is to promote pattern-noticing 558.

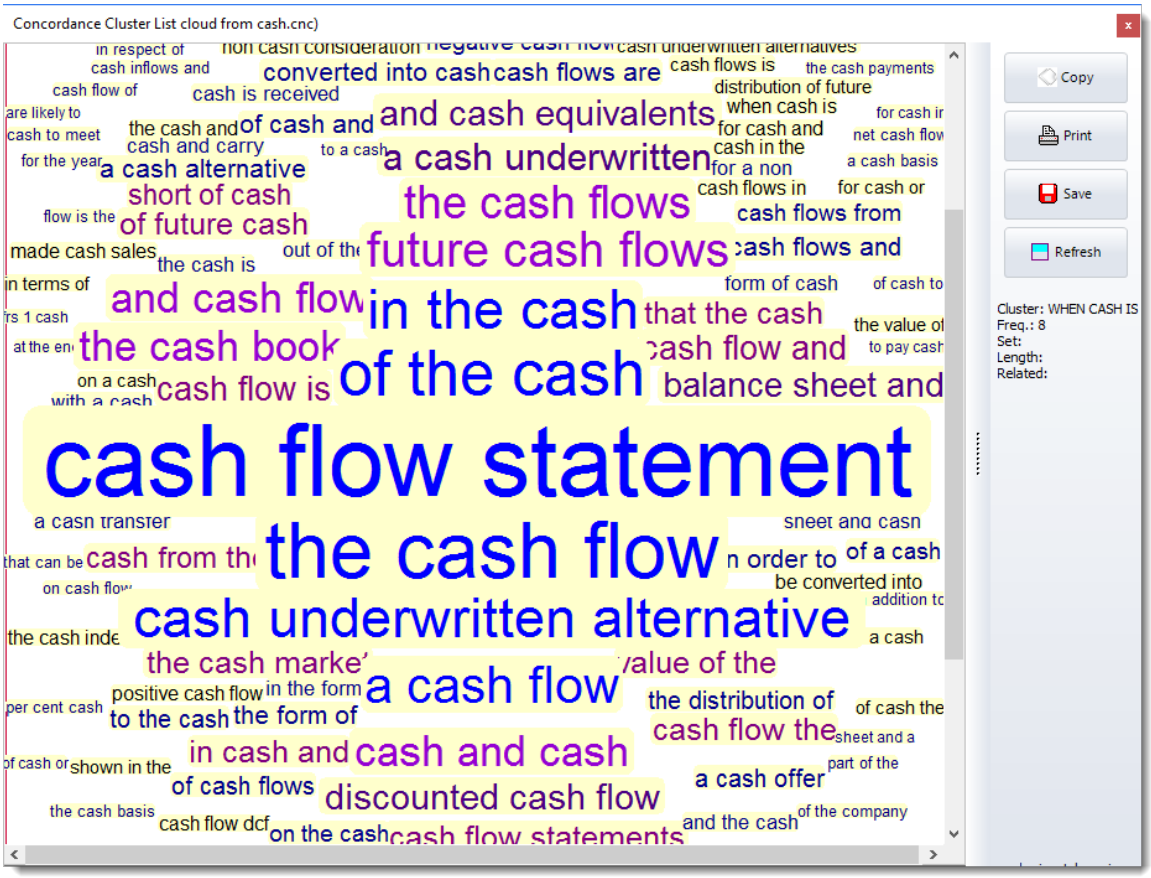
How to get here

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

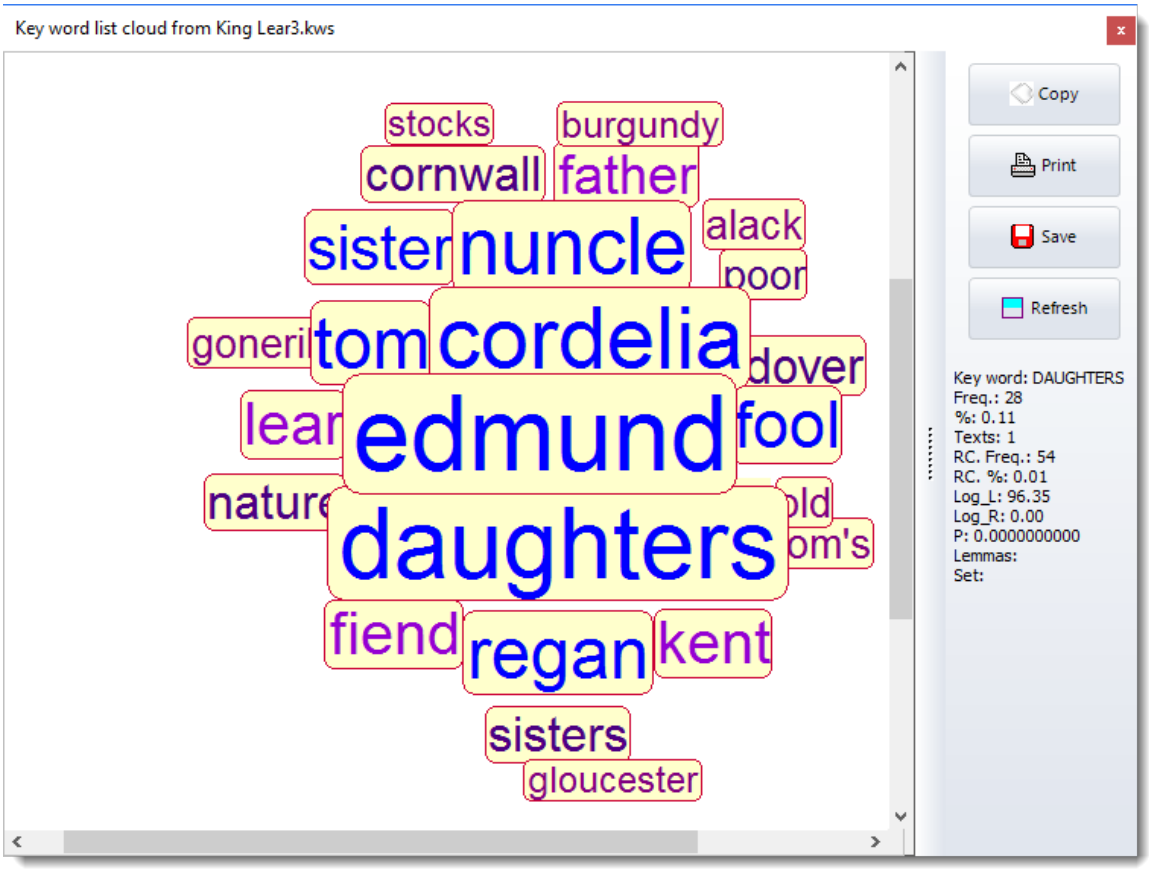
Examples

In this case of collocates of 219 ***cash*** 219 you can get a word cloud based on any column of data.

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



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



In this case the word cloud was based on the key words of the Shakespeare play, *King Lear*. Details of the highlighted word **daughters** are shown to the right.

What you can see and do

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

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

You can change the word cloud settings in the main Colours setting 114 in the Controller. To force lower-case, Title case or UPPER case for the display, set the Layout 60 settings first.

The font sizes reflect the range of values in your data. The font 118 is the one you may choose for any of your standard displays.

There are some settings including *minimum font size* you can play with, visible at the bottom right of the (resizeable) word cloud window.

5.29 zap unwanted lines

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

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

See also : recompute lists of file-names 242, reduce data to N entries 71.

Tags and Markup

6 Tags and Markup

What is markup for?

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

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

You may want to translate HTML or SGML 543 tags or entity references: if your text has **É** you probably want to see **É**.

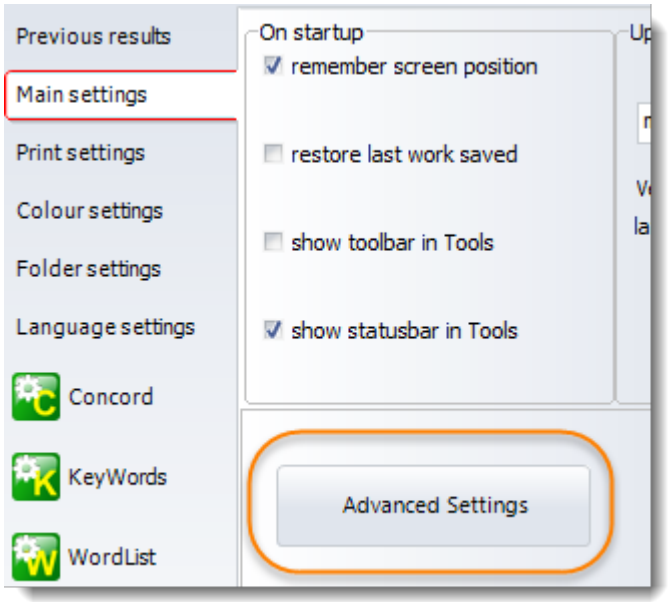
You may wish to select within text files, e.g. cutting out a header or getting only the conclusions, instead of using the whole text.

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

You can see the effect of choosing tags if you select the Choose Texts option, then press the View 401 button. Any retained tags will be visible, and ignored tags replaced by spaces.

Tags and Markup Settings

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



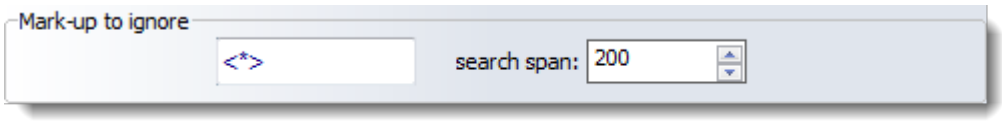
See also: Guide to handling the BNC, Handling Tags 158, Making a Tag File 172, Showing Nearest Tags in Concord 223, Concord Sound and Video 258, Tag Concordancing 221, Types of Tag 159, Viewing the Tags 401, Using Tags as Text Selectors 167, Tags in WordList 367, XML text 189

6.1 choices in handling tags

ignore all tags

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

example :



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

ignore some tags and retain others

If you want to ignore some but retain others, you will need to prepare a tag file 172 which lists all those you want to keep. These will then appear in your word lists and concordances.

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

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

Tags to retain will only be active if there's a file name visible and you have pressed the *Load* or *Clear* button. If you press *Load*, you will see which tags have been read in from the tag file. Any tags to retain get saved with the words but can later be shown or hidden as you prefer. Here are examples 186.

translate entity references into other characters

If you use XML, SGML or HTML 543 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 172 which contains the strings you want to translate. Then choose *Main Settings | Advanced | Tags & Markup | Entity File (entities to be translated)* and choose your entity file. WordSmith will then translate any entity references in this file into the corresponding characters.

to load up these tag files automatically

See Custom Settings 165.

See also: Guide to handling the BNC, Overview of Tags 157, Making a Tag File 172, Showing Nearest Tags in Concord 223, Tag Concordancing 221, Types of Tag 159, Viewing the Tags 401, Using Tags as Text Selectors 167, Tags in WordList 374

6.2 tag-types

Mark-up to ignore

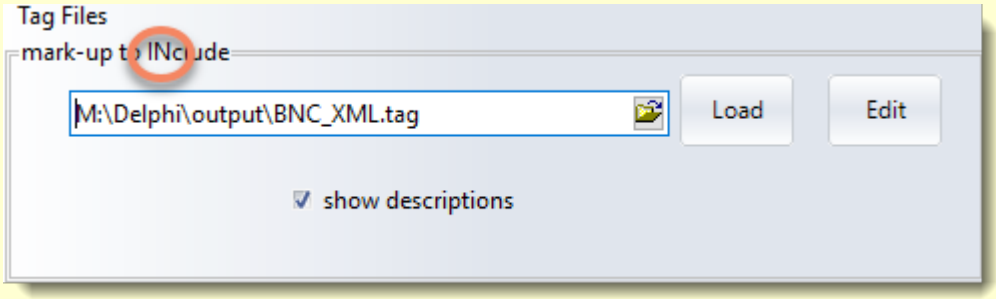
The Tags to Ignore 158 option allows you to specify tags which you do not want to see in the concordance or word list results.

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.)

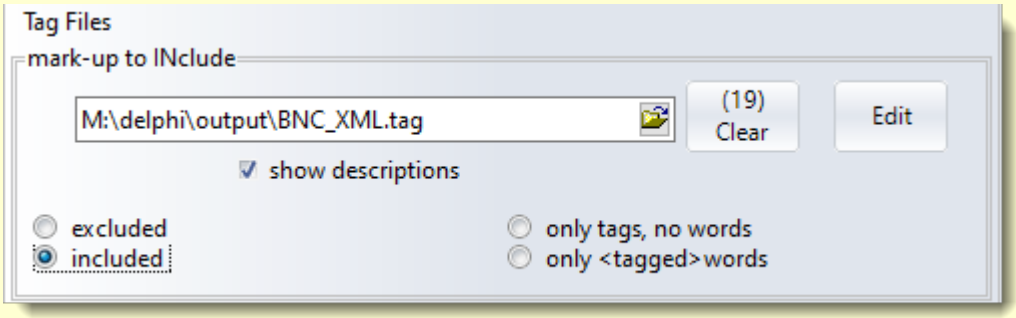
Custom settings

Lets you specify a set of settings you will want to load up all in one go, e.g. when WordSmith starts up.
You can specify various alternatives for different types of work, on plain text, British National Corpus, Shakespeare text. You can edit the options 165 and their effects.

Mark-up to INclude



The Tags to be INcluded 172 option allows you to specify a tag file, containing tags which you do want to see in the concordance or word list results.



Here 19 tag file elements have been loaded. The user will make word lists where these tags are included. See examples 186.

Mark-up to EXclude

The Tags to be EXcluded 172 option allows you to specify a different tag file, containing stretches of tags which you want to find and remove in the concordance or word list results.

Entity file

HTML, XML and SGML 543 use so-called entity references for symbols which are outside the standard alphabet, e.g. **ét&eacute** which represents **été**.

The Tags to be Translated 158 option allows you to specify entity references which you want to convert on the fly, such as **&eacute**.

Document header

Document Header

You can specify a document header to be routinely cut out: standard text at the beginning of each file in a corpus. For example, each BNC text file has a long section starting with **<teiHeader>** and ending with **</teiHeader>** . Or a copyright notice.

The process cuts by looking for the *Document header ends* mark-up and deleting all text to that point.

Only if containing

Helps select texts which contain words or phrases you specify.

Only part of file

Helps you process only certain sections of the text files, such as introductions or conclusions.

Media tags

multimedia markers

Text files can be tagged for reference to sound or video files which you can hear or see. For example, a text might contain something like this: **blah blah blah ...<a href=http://gandalf.hit.uib.no/c/l/32401-1.mp3> blah blah** etc. A concordance on **blah blah** could pick up the tag so you can hear the source mp3 file. See defining multimedia tags 162.

Tag format

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 165, sections 165, sentences 164 and paragraphs 165. Type these in the appropriate spaces when selecting Text Characteristics 145.

tags within 2 separators 534

These tags are often used to signal the part of speech of each word; they're also widely used in HTML, XML, SGML 543 for "switches", e.g. **<H1>** to switch on Heading 1 style and **</H1>** to switch it off again. You should use the same opening and closing symbols, usually some kind of brackets, for all your tags (as the British National Corpus does using SGML or XML 543 markup): **<Noun>,<Verb>,<Pronoun>**.

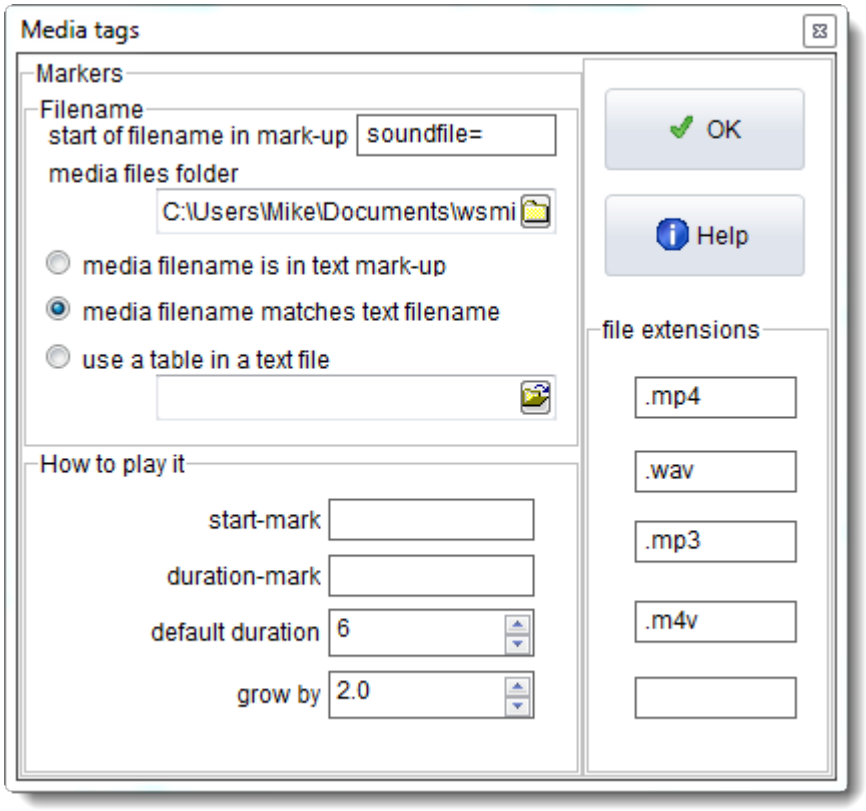
(A particular sub-variety of tags within 2 separators sometimes used is tags with underscores at the left and space at the right like this

To process these, you will need to declare the underscore a valid character 146, or else (recommended) convert your corpus 492 to a format like.

See also: Overview of Tags 157, Handling Tags 158, Making a Tag File 172, Showing Nearest Tags in Concord 223, Tag Concordancing 221, Viewing the Tags 401, Using Tags as Text Selectors 167, Concord Sound and Video 258, Guide to handling the BNC.

6.2.1 multimedia tags

In this screenshot you see an example of how to define your multimedia tags. This is accessed from *Main Settings* | *Advanced* | *Tags* | *Media Tags*.



File Extensions

The file extensions (**.wav, .mp3, .avi,** etc.) define the file types which your computer can play. Of course this function does require your computer to be able to handle sound or video if it is to work -- Windows uses the file extension to know how to play it.

Video files will require the free VLC Media Player to be installed (see http://www.vlcapp.com/).

Filename

The sound or video file-name might be

1. specified in a tag
2. the same name as the text file-name but with another extension such as **.wav**
3. found in the tag and interpreted using a table you have created previously. To do this, make each line like this:

<s1>=c:\my_corpus_sounds\angry_man.wav 560 2
<s2>=c:\my_corpus_sounds\happy_little_girl.mp3 980 2

where each line has the tag found in the text file, followed by **=** then the desired value.

If it is in the tag mark-up, to process a reference like **<a href=http://gandalf.hit.uib.no/c/l/32401-1.mp3>** in the source text, the **=** character is sufficient to define where the start of the filename begins. In this case, what follows **=** is a web address. For a text containing tags like this **<sound$$C:\mysounds\talk.wav>**, you'd put **$$** to show the start of filename. For the concordance example 258, **soundfile=** is adequate to identify where the filename begins.

The *media files folder* will be needed (for cases 1 and 2 above) if the sound files are not stored in the same folder as your text files.

How to play it

start-mark, duration-mark (optional)

You can indicate markers for start and duration if necessary. They would be needed if your tag contained e.g.

<a href=http://gandalf.hit.uib.no/c/l/32401-1.mp3 start=0360 play=5> or if the sound file is the same as the text file with a different extension.

If so, you'd specify *duration-mark* as **play=** and *start-mark* as **start=** (because that is how they are marked in your text files)

Times are measured in seconds.

duration

You can specify a default duration as in the screenshot: 6 seconds. More may be needed especially if the sound tags are not spaced closely together in the text file.

If no start or duration indication is given, the whole sound or video file will be played.

If there are no duration and start position markers, the first number will be interpreted as start position and the second as duration, so a tag like this: **<sound$$C:\mysounds\talk.wav 15 5>** in your text file means "play **c:\mysounds\talk.wav** starting 15 seconds from the beginning and play for 5 seconds". If there's only one number as in **<sound$$C:\mysounds\talk.wav 15>**, that means "play **c:\mysounds\talk.wav** starting 15 seconds from the beginning and play for the default number of seconds".

defaults

The defaults are: play **.mp4**, **.mp3** and **.wav** files. Once you've completed this, save your defaults 110 for next time.

See also: Sound and Video in Concord 258, Obtaining Sound and Video files 263, Overview of Tags 157, Making a Tag File 172, Tag Handling 158, Tag Concordancing 221, Showing Nearest Tags in Concord 223, Viewing the Tags 401, Types of Tag 159, Guide to handling the BNC

6.2.2 start and end of text segments

WordSmith attempts to recognise 4 types of text segment: sentences, paragraphs, headings, sections. Processing is case sensitive. You can use **<Enter>** and **<Tab>** as strings representing an end of paragraph or a tab in your texts. For sentence ends, **auto** is another option.

Define these in your language settings 77.

Sentences

For example, **<s>** might represent the beginning of a sentence and **</s>** the end. If you leave the choice as **auto**, ends of sentences are determined by according to the definition of a sentence 533

which gives a good approximation. (There is no 100% accurate way 401 of handling sentence recognition.)

Paragraphs

For example, **<p *>** or ** ** might represent the beginning of a paragraph and ** ** 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>, ** etc. is encountered. If there are no closing ** ** tags in the entire text then paragraphs will be counted each time the opening paragraph tag is found.

See also: Overview of Tags 157, Handling Tags 158, Showing Nearest Tags in Concord 223, Tag Concordancing 221, Types of Tag 159, Viewing the Tags 401, Using Tags as Text Selectors 167, Guide to handling the BNC.

6.3 custom settings

Custom Tagsets

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

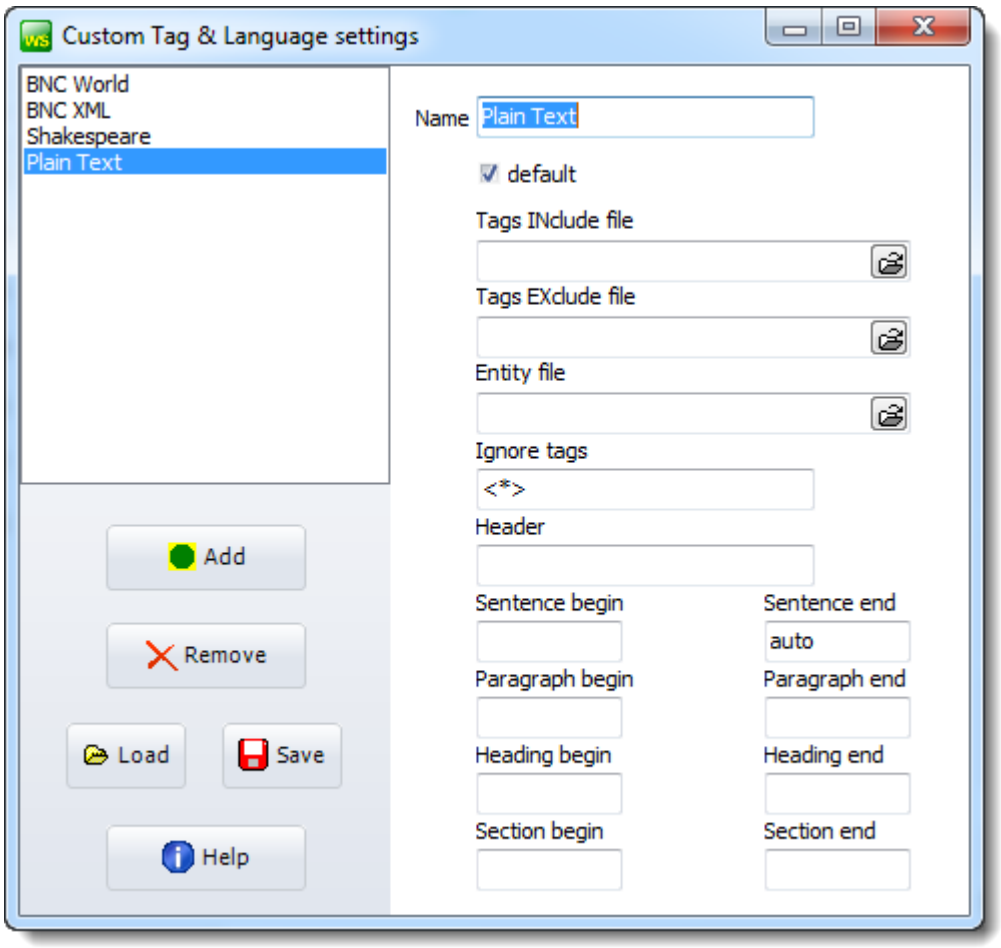
The point of it...

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

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

How to do it

Press the Edit button to create or edit the custom settings (the file is called **`custom_tag_settings.xml`** and it'll get saved in your **`Documents\wsmith7`** folder).



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

Fill in the boxes and press Save.

- All boxes will have leading and trailing spaces removed.
- Use **`auto`** for automatic processing e.g. of sentence ends.
- Checking the *default* box means that this set gets chosen by default and any tag or entity files will get automatically loaded for you.
- Don't forget to save. These settings will load automatically.

See also : Tags as text selectors 167

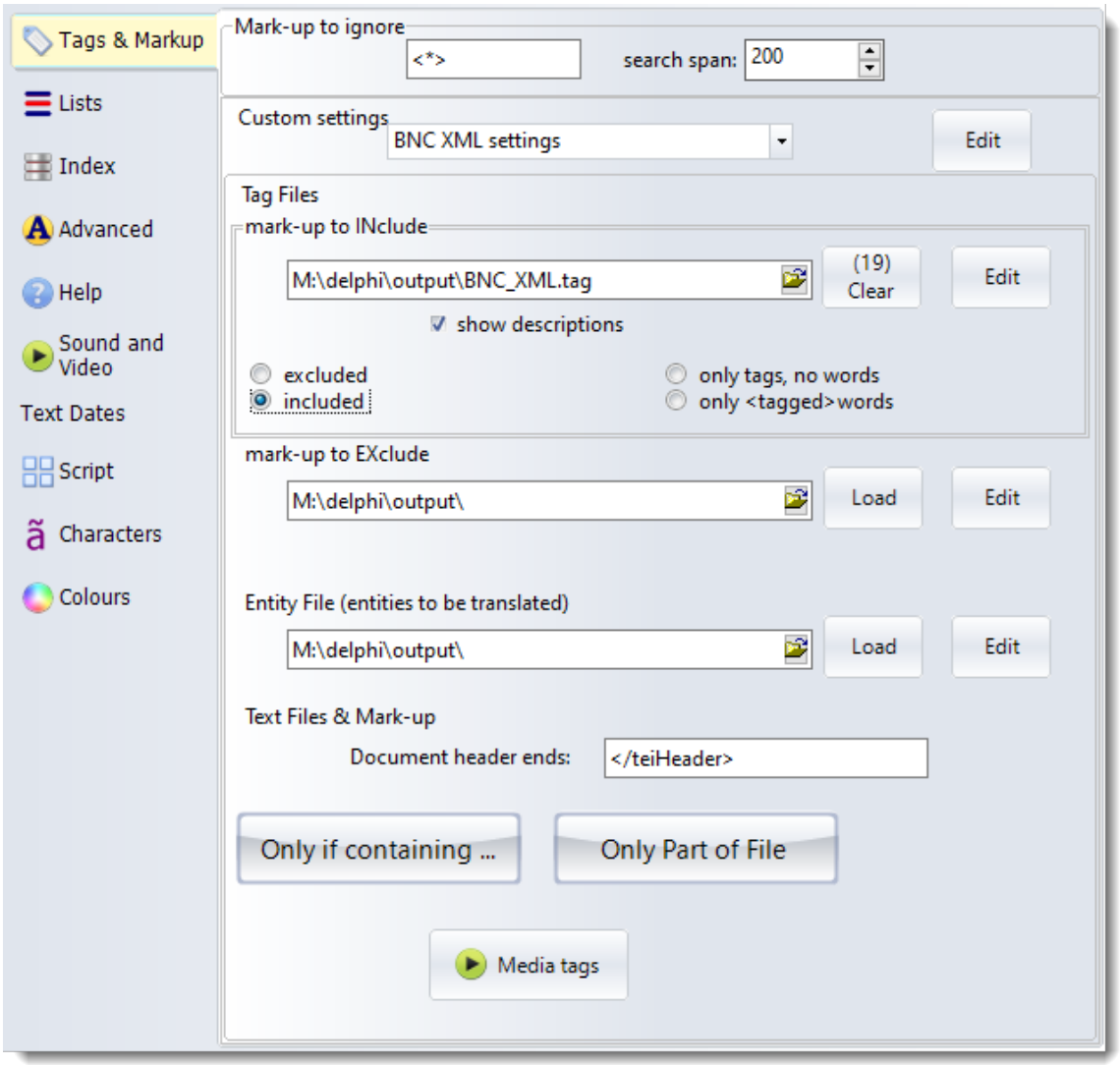
6.4 tags as selectors

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

These functions are available from *Main Settings* | *Advanced* | *Tags* | *Only If Containing* 171 *or* *Only Part of File* 168.

Defaults

The defaults are: select *all* sections of *all* texts selected in Choose Texts 52 but cut out all angle-bracketed tags.



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 171 to check whether it contains the words you've specified;
2. Cut out any section specified as a "section to cut 168";
3. If there are "sections to keep 168", cut out everything which is not within them;
4. Cut start of each line 168, if applicable;
5. Process any entity references you want to translate 158;
6. Ignore 158 any tags not to be retained or excluded (see the "Mark-up to ignore" section of the screenshot above).

See also: Overview of Tags 157, Making a Tag File 172, Tag Handling 158, Tag Concordancing 221, Showing Nearest Tags in Concord 223, Viewing the Tags 401, Types of Tag 159, Guide to handling the BNC, XML text 189

6.4.1 tags to select between texts

Text Files and Mark-up

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

These functions are available from *Main Settings* | *Advanced* | *Tags* | *Only If Containing* 171 *or* *Only Part of File* 168.

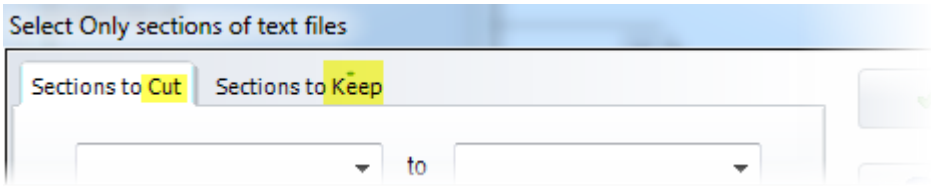
6.4.1.1 part of file:selecting within texts

The point of it

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

Cut out or Keep?

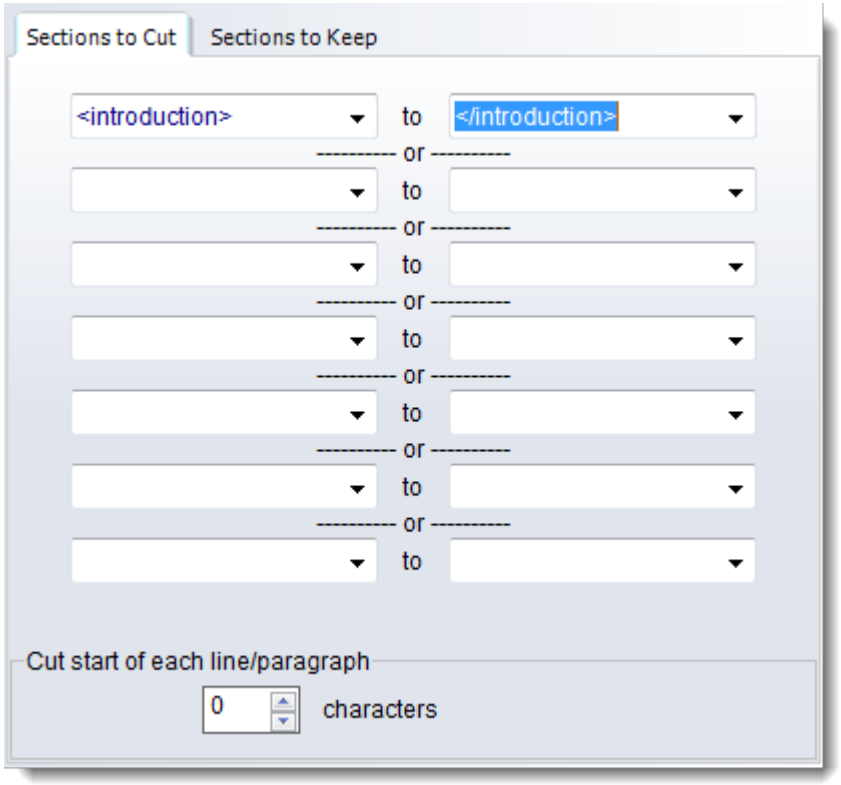
Press the *Cut* or *Keep* tab to choose to cut out certain sections, and/or only to use certain sections.



Sections to Cut

Note: see Document Header 161 to remove a document header such as **</header>**.

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



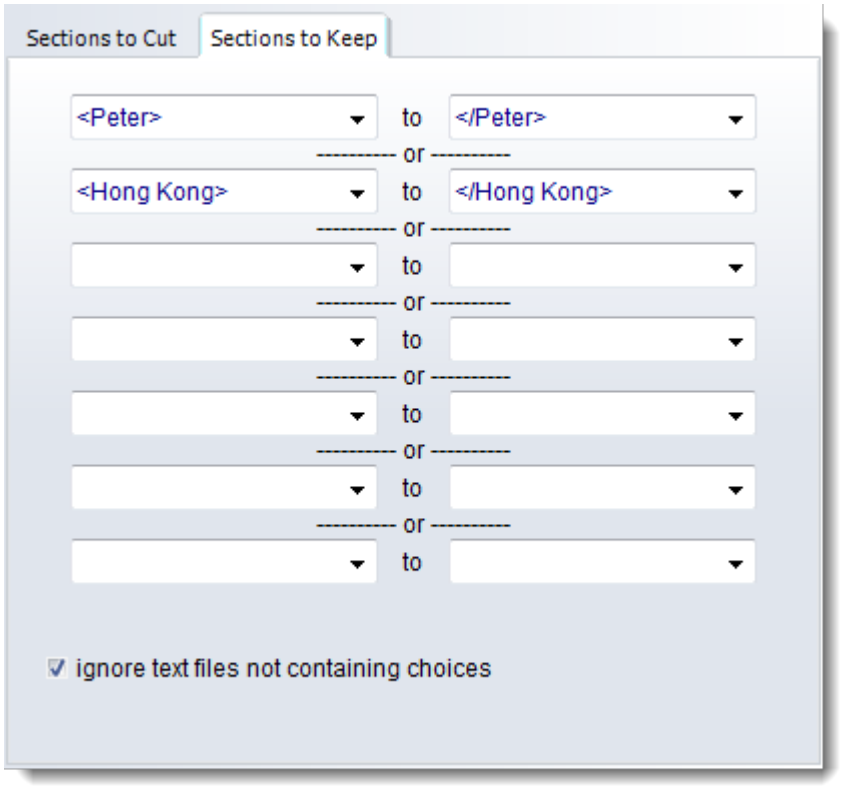
Cut start of each line/paragraph

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

Sections to Keep (contexts)

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

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



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

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

If you used **<H1>** to **</H1>** with this function in HTML 543 text you'd get all the major headings in your texts, however many, but nothing else.

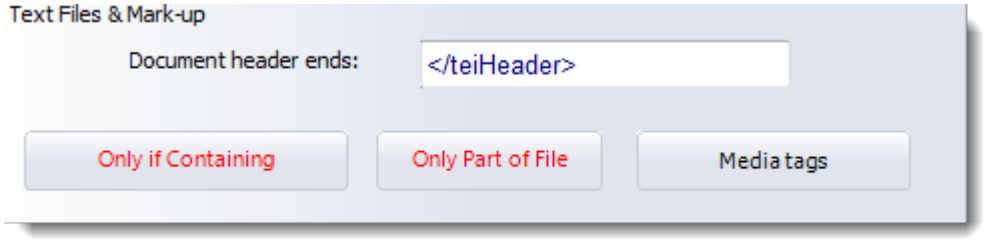
The "off" switch doesn't have to look like the "on" switch -- you could keep, for example, **<INTRO>** to **</BODY>** and thereby cut out the conclusion if that comes after the **</BODY>**.

Ignore text files not containing choices

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

OK

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



See also: Tags as Selectors 167, Only if containing <x> 171, Guide to handling the BNC.

6.4.1.2 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.

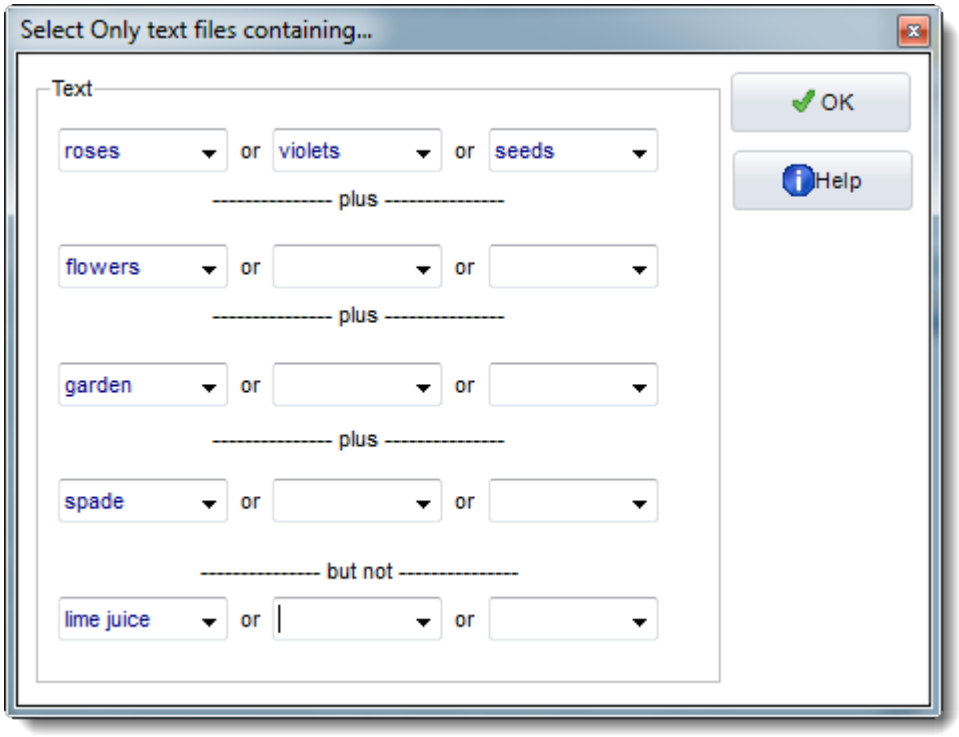
Knowing that the BNC uses **stext>** for spoken texts, **sex=m** for males, **age=5** for speakers aged 60 or more, you can get WordSmith to filter your text selection. It will search through the whole of every text file (not just the tags or header sections, in fact the first 2 megabytes 544 of the file) to check that it meets your requirements.

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

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

After your text files have been processed, you will be able to see which met your requirements in the Text File choose window 52 and can save the list for later use as favourites 59.

Examples:



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

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

If you want to know which text files met your criteria and which did not you can choose save text file data in the Advanced options.

See also: Tags as Selectors 167, Selecting within texts 168, Extracting text sections 478, Filtering your text files 479 using Text Converter 471, Guide to handling the BNC

6.5 making a tag file

Tags for your text

Each tag you define has to match the structure of your text. If it doesn't match you won't get the results you want.

detailed explanation

If your corpus is formatted like this

```
some_DD parts_NN2 of_IO Britain_NP1
```

then you may wish to get it re-structured (easily & quickly) like this

```
<DD>some <NN2>parts <IO>of <NP1>Britain
```

using the Text Converter 492.

If your corpus has lots of mark-up within each < > like this

```
<w c5="VVG-NN1" hw="shiver" pos="VERB">shivering </w><w c5="VBZ"
hw="be" pos="VERB">is </w>
<w c5="AT0" hw="a" pos="ART">a </w><w c5="NN1" hw="body"
pos="SUBST">body </w>
<w c5="NN1" hw="reaction" pos="SUBST">reaction
```

you need to know how many attributes there are for each piece of mark-up. In this case (BNC XML) we have a four: **w** then an attribute beginning **c5** and ending **"** then lemma information beginning **hw**, and a **pos** section. Each attribute can be handled with an asterisk. So you might have to tell WordSmith to find **<w c5="NN1" * *>***. That means, "*find me a stretch beginning <w c5="NN1" with any two following attributes, then a closing > and do not require a word-break after it*." This is explained also in the XML text section 189.

Tag Syntax

Each tag is case sensitive.

Tags conventionally begin with **<** and end with **>.** You will be asked to confirm any tag with different first & last characters.

The ideal format for use with WordSmith, for single-word mark-up, is **<TAG>word**. In other words your tag immediately precedes the word it refers to. Tag as prefix to word. For longer stretches, **<OPENING_TAG>blah blah blah</OPENING_TAG>**. In other words a tag is used to signal the start of mark-up, and **/** plus the same tag is used to signal the end.

You can use

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

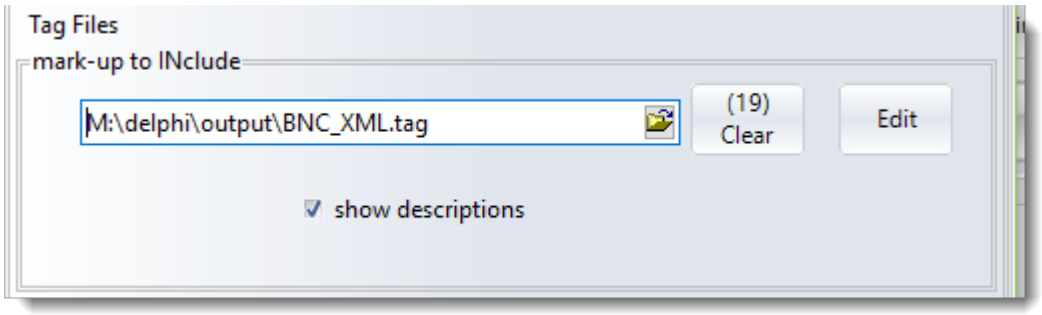
Don't use **[** to insert comments in a tag file, since **[** is useful as a potential tag symbol. You can use # to represent a number (e.g. **<h#>** will pick up **<h5>, <h1>**, etc.). And use **?** to represent any single character (**<?>** will pick up **<s>, <p>**, etc.), or * to represent any number 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 141.

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

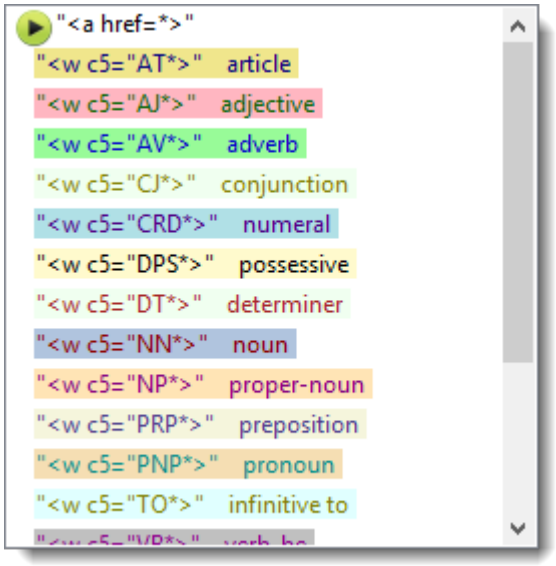
Any number of pre-defined tags can be stored. But the more you use, the more work WordSmith has to do, of course and it will take time & memory ...

Mark-up to INclude

A tag file for tags to retain contains a simple list of all the tags you want to retain. Sample tag list files for BNC handling (e.g **bnc_XML.tag**) are included with your installation (in your **Documents\wsmith7** folder): you could make a new tag file by reading one of them in, altering it, and saving it under a new name.



Here a tag file has been loaded; it had 19 entries.



and here is a word list based on BNC spoken classroom data where the tag file was used:

WordList

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	% Dispersion	Lemmas	Set
1	<pronoun>YOU	15,444	3.75	58	100.00	0.97	<pronoun>
2	<article>THE	15,035	3.65	58	100.00	0.95	<article>
3	<pronoun>IT	13,173	3.20	57	98.28	0.96	<pronoun>
4	<conjunction>AND	10,363	2.52	58	100.00	0.98	<conjunction>
5	<article>A	9,666	2.35	58	100.00	0.97	<article>
6	<determiner>THAT	7,527	1.83	58	100.00	0.96	<determiner>
7	<pronoun>I	7,371	1.79	58	100.00	0.97	<pronoun>
8	OF	7,338	1.78	57	98.28	0.97	
9	<infinitive to>TO	6,009	1.46	58	100.00	0.96	<infinitive to>
10	<verb-be>IS	5,513	1.34	58	100.00	0.96	<verb-be>
11	<adverb>SO	4,971	1.21	58	100.00	0.95	<adverb>
12	<determiner>WHAT	4,944	1.20	57	98.28	0.98	<determiner>
13	<verb-do>DO	4,583	1.11	58	100.00	0.96	<verb-do>
14	<pronoun>WE	4,375	1.06	57	98.28	0.94	<pronoun>
15	<numeral>ONE	4,290	1.04	57	98.28	0.92	<numeral>
16	YEAH	3,987	0.97	56	96.55	0.85	
17	<preposition>IN	3,965	0.96	58	100.00	0.93	<preposition>
18	ERM	3,480	0.85	57	98.28	0.93	

Here is a concordance using the Noun tag:

N	Concordance	Set	Tag
1	and <adjective>entrepreneurial <noun>**abilities.** <conjunction>And		<noun>
2	as there <verb-be>are <article>no <noun>**absolutes** erm <pronoun>you		<noun>
3	have <adverb>far <determiner>more <noun>**accidents** er <preposition>in		<noun>
4	to>to <verb-have>have <noun>bank <noun>**accounts** <preposition>in		<noun>
5	<modal>**will,** erm <article>a <noun>lot of <noun>**accusations,** none <verb-have>have		<noun>
6	control <determiner>these <noun>kind of <noun>**activities.** <preposition>As <article>		<noun>
7	<proper-noun>Koresh's <adjective>dubious <noun>**activities** <pronoun>it <verb-be>		<noun>
8	is er <conjunction>But <determiner>more <noun>**adults** <verb-be>are <adverb>now		<noun>
9	. <proper-noun>Cautioning <verb-have>has <noun>**advantages** <conjunction>and not		<noun>
10	<verb>recruiting <proper-noun>Manchester <noun>**Adventists** <verb>using		<noun>
11	legal <conjunction>and <adjective>medical <noun>**advisors.** <pronoun>They		<noun>
12	about <possessive>his <noun>love <noun>**affairs** <conjunction>and <noun>		<noun>
13	, <article>every <noun>day, <noun>enquiry <noun>**agents** <verb-be>are <verb>used		<noun>
14	was, <initials>F <initials>B <initials>I <noun>**agents** who <verb-be>were <verb>		<noun>
15	a <numeral>hundred <adjective>armed <noun>**agents** <verb>set <adverb>out		<noun>
16	we <verb>got <numeral>five <noun>tracing <noun>**agents** <infinitive to>to <verb>come		<noun>
17	<adjective>big <adjective>international <noun>**airlines,** <adverb>especially		<noun>
18	**er** of <article>the er <adjective>European <noun>**airlines** <preposition>as <article>a		<noun>
19	<numeral>two <adjective>scheduled <noun>**airlines** <preposition>in <article>the		<noun>

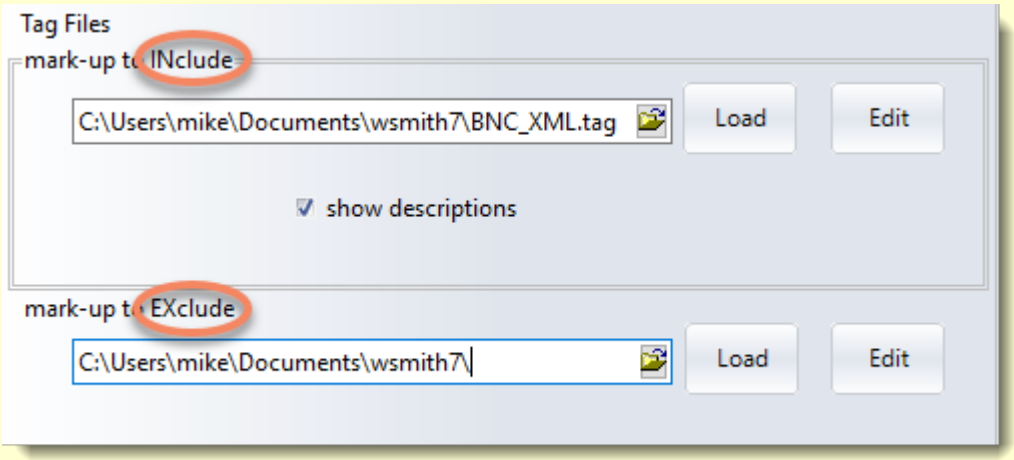
… easy to see adjectives come before nouns…

You can hide the tags and keep the colours using *View* | *Tag hiding* | *Tagfile tags*.

if <tag>WORD

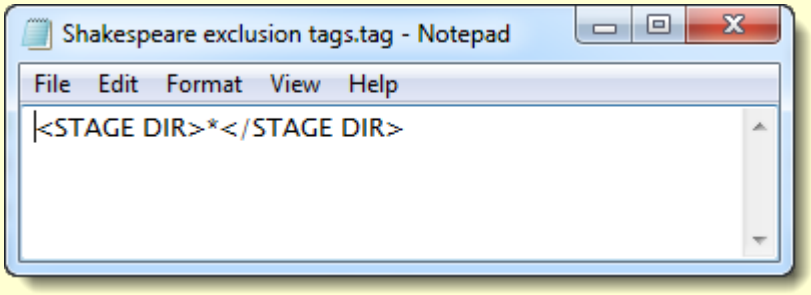
... this means, if a tag is found immediately to the left of a word (such as **<VERB>wanting** for example), WordSmith can by default treat the **<VERB>** tag as a tag, or can ignore the word, only recording the tag. This is useful if you do not want to keep the words but are only interested in their tags.

Mark-up to EXclude



A tag file for stretches of mark-up like this **<SCENE>A public library in London. A bald-headed man is sitting reading the News of the World.</SCENE>** where you want to exclude the whole stretch above from your concordance or word list, e.g. because you're processing a play and want only the actors' words. Mark-up to exclude will cut out the whole string from the opening to the closing tag inclusive.

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



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

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

Legal syntax examples would be:

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

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

<SCENE>*
</SCENE>

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

<SCENE>*<\SCENE>

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

With your installation you will find (**Documents\wsmith7\sample_lemma_exclude_tag.tag**) included, which cuts out lemmas if constructed on the pattern **<lemma tag="*>*</lemma>**, i.e. with the word **tag**, an equals sign and a double-quote symbol, regardless of what is in the double-quotes.

Colours

Tags will by default be displayed in a standard tag colour 114 (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 can be seen in the list of colours in the colours tab.

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

Sound and Video

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 162. 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 258.

Descriptive Label

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

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

Show descriptions

If you define a description, WordSmith will by default show that instead of the formal tag. The "Show descriptions" check box switches on or off the renaming of your tags to the format in the description section you define.

Section Tag

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

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

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

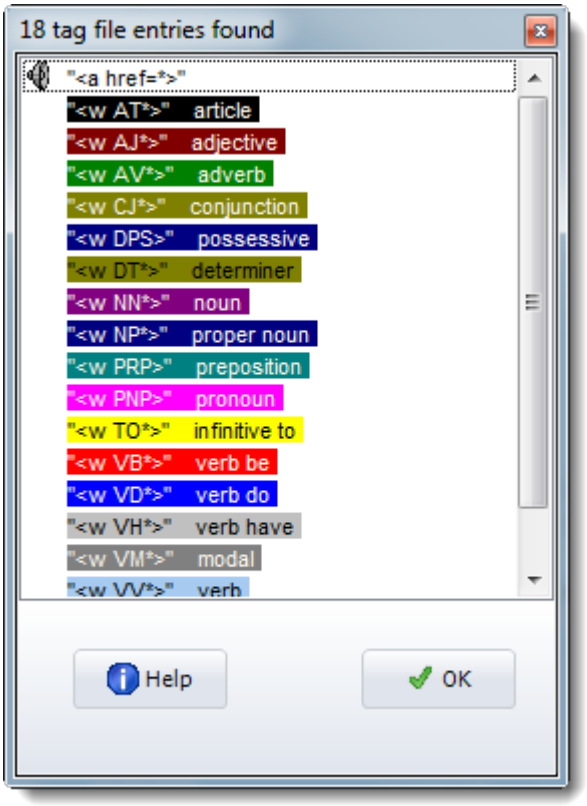
Tagstring_only tags

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

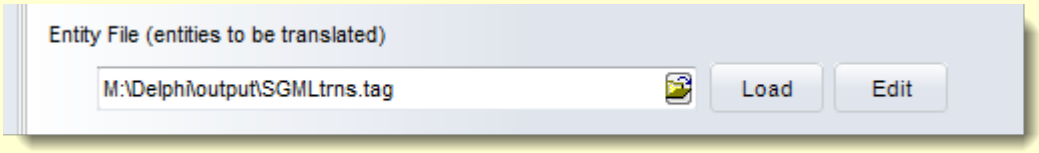
```
<title> /tagstring_only
</title> /tagstring_only
```

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

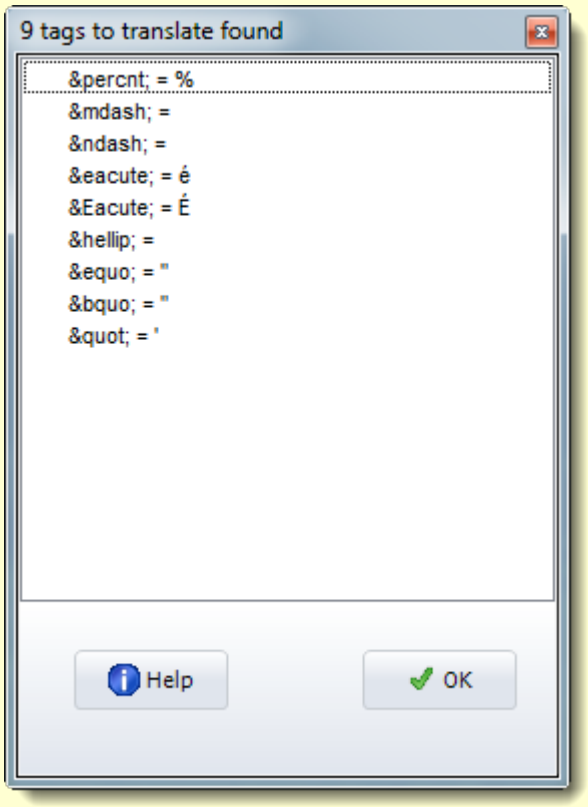
Here is an example of what you see after selecting a tag file and pressing "Load". The first tag is a "play media" tag, as is shown by the icon. You can see the cream on purple colour for nouns too. The tag file (**BNC World.tag**) is included in your installation.



Entity File (entities to be translated)



If you load it you might see something like this:



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

```
&Eacute; É
&eacute; é
```

In the screenshot above, the sample tag file for translation (**Documents\wsmith7\sgmltrns.tag**) which is included with your installation has been loaded. You could make a new one by reading it in, altering it, and saving it under a new name.

See also: Overview of Tags 157, Handling Tags 158, Showing Nearest Tags in Concord 223, Tag Concordancing 221, Types of Tag 159, Viewing the Tags 401, Using Tags as Text Selectors 167, Guide to handling the BNC.

6.6 modify source texts

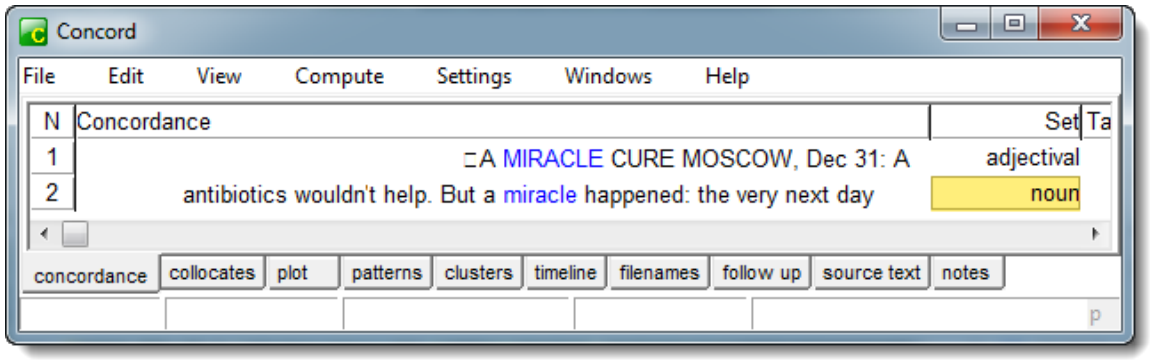
The point of it...

This function enables you to modify your original source texts as a result of concordance work you've done. In this way, your work can get saved in the source texts themselves. For example, you might want to save user-defined categories, or search-phrase results where you have decided a phrase is a multi-word unit.

Note: this procedure does alter your source texts. Before each is altered for the very first time, it is backed up (original filename with **.original** extension) but any change to your source texts or corpora must be done with caution!

User-defined categories

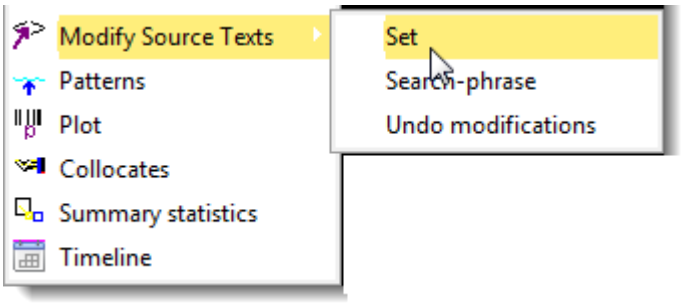
For example, suppose you have marked your concordance lines' Set column 43 like this:



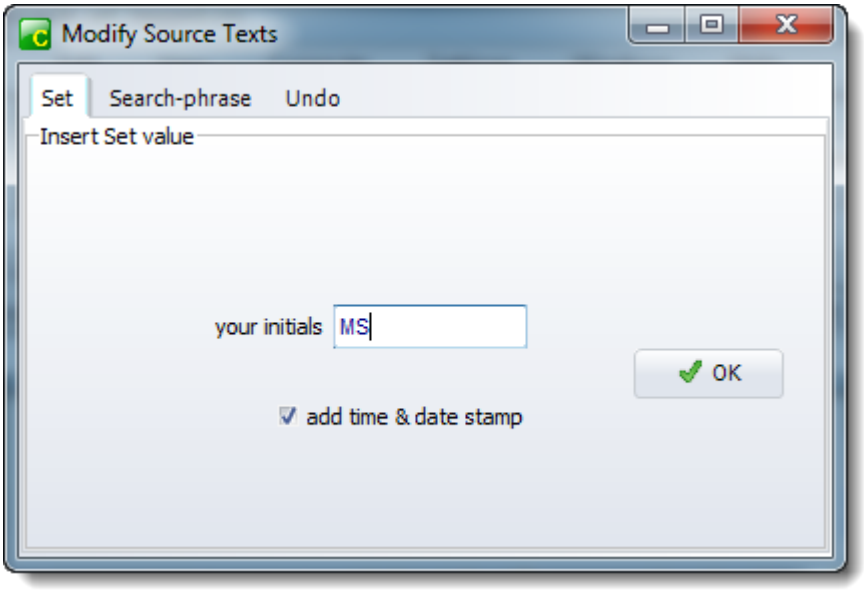
where the first line with **miracle** pre-modifies the noun **cure** and is marked a**djectival** but the second is an ordinary noun, and wish to save this in your original source text files.

How to do it

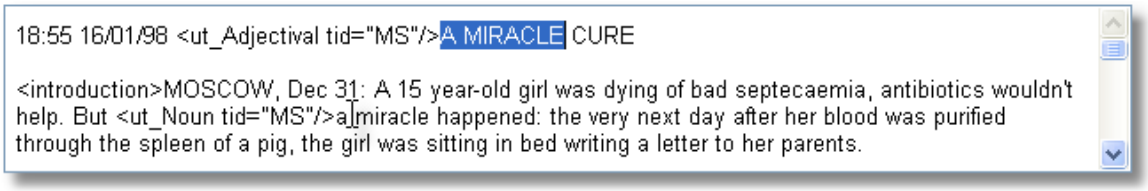
Choose *Compute* | *Modify Source Texts*.



and if you want to save the Set choices, choose OK here:



and the set choices will be marked as in this example:

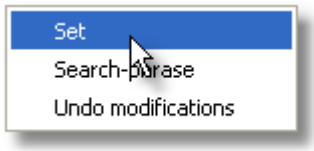


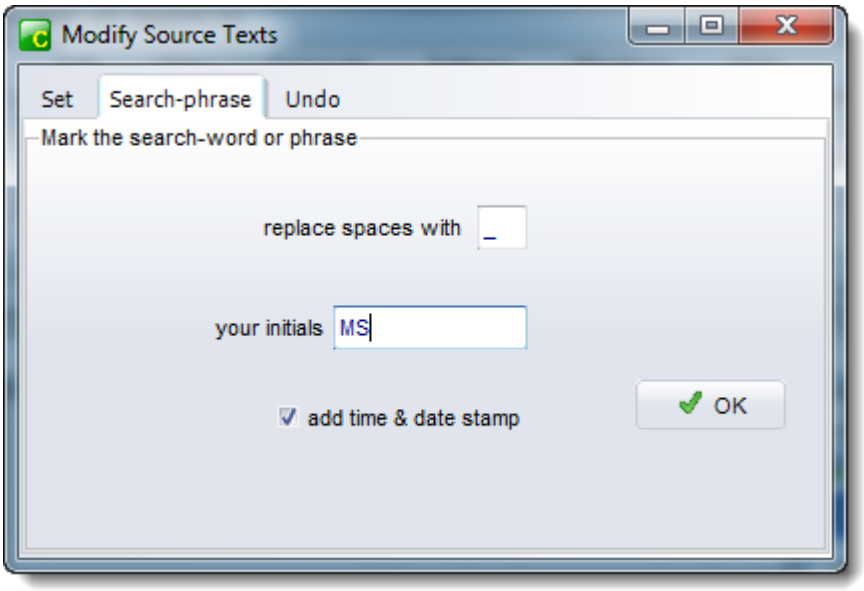
(seen by double-clicking the concordance line to show the source text).

Multi-word unit search phrase

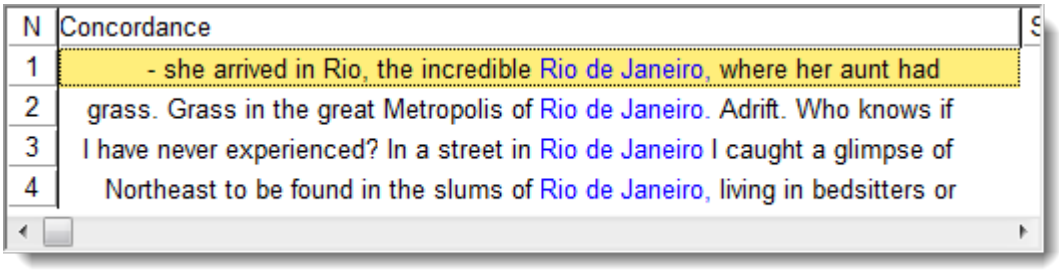
Alternatively if you choose the search-phrase option:

and

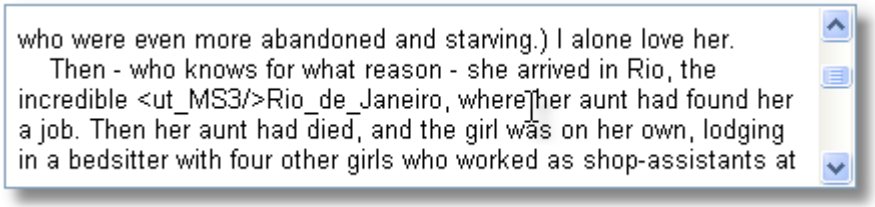




then any search word containing a space will have underscores (or whatever other character you choose above) in it to establish multi-word units:



Here, the search word or phrase was **`Rio de Janeiro`**, and the result of modifying the source texts was this:



Add Time & Date stamp option

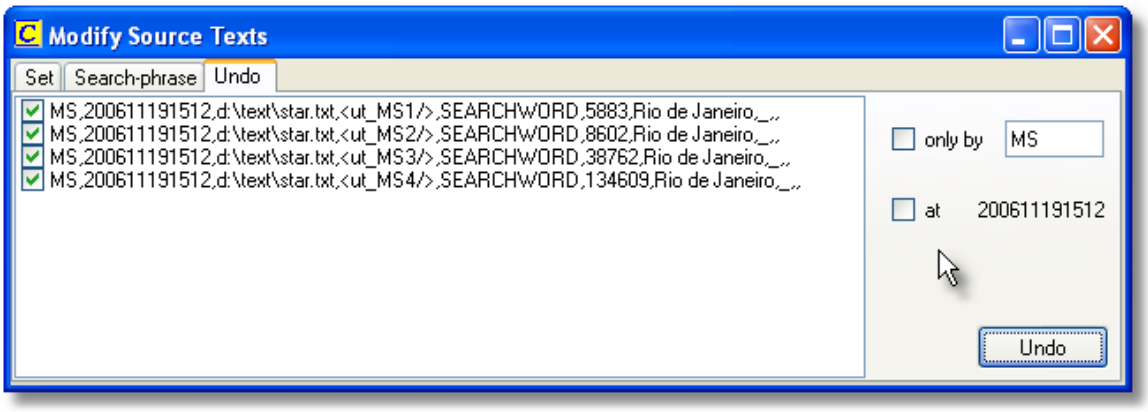
This keeps a log of all your changes, enabling the changes to be undone later.

Initials option

Adds your initials to the changes. Leave empty if not wanted. The **<ut_MS3/>** tag above means a user whose initials were **MS** made this change and it was the 3rd change.

To undo previous changes

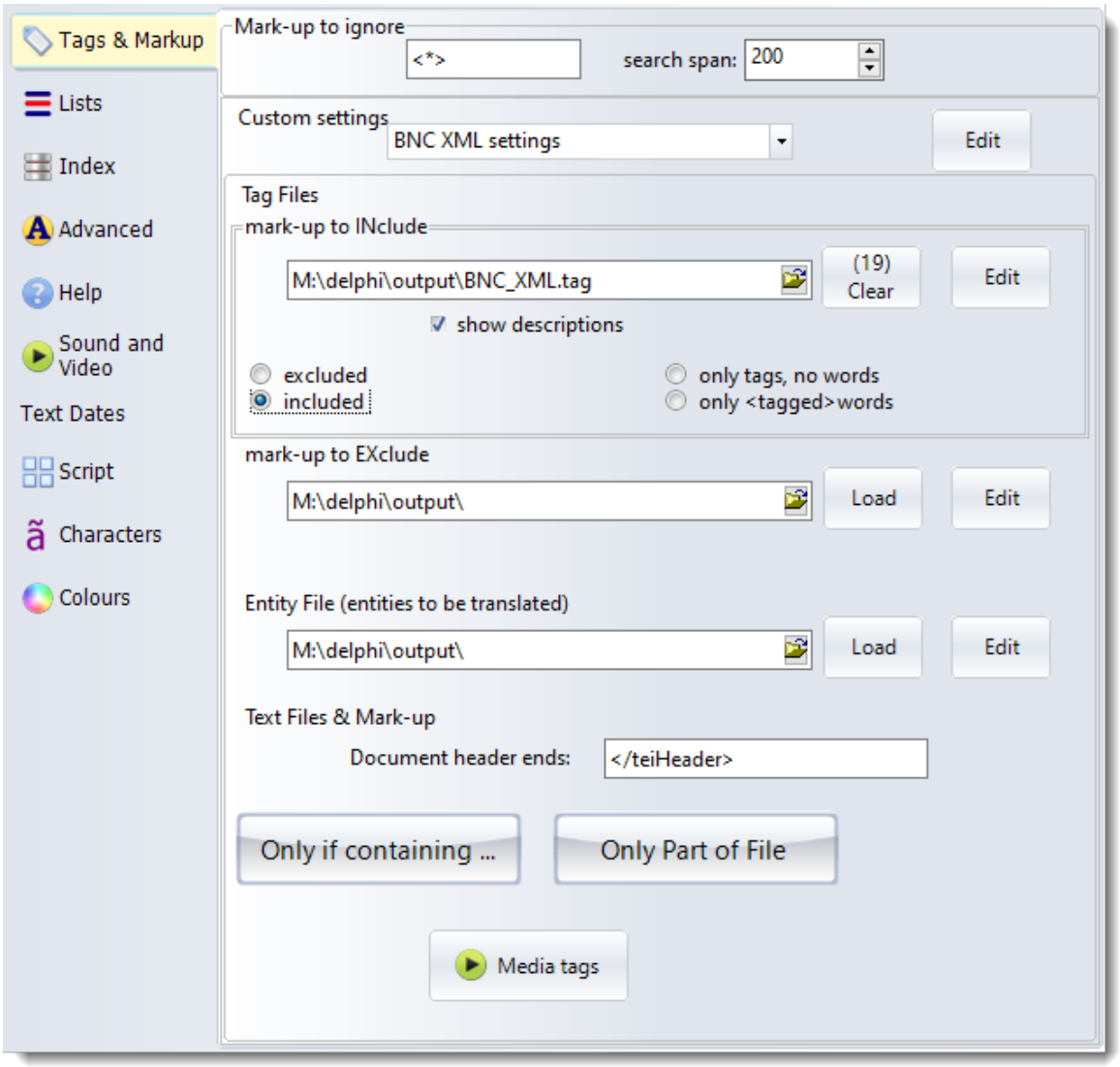
If you have used the "time and date stamp" option shown above, you will be able to undo the modifications. The undo window shows all your log. You can choose all those done on a certain day, or by the person whose initials are visible at the right. Here we see the 4 modifications changing **Rio de Janeiro** into **Rio_de_Janeiro**.



See also: adding markup when exporting a concordance to TXT, .XLS etc. 245 user-defined categories 43

6.7 examples

The mark-up to INclude section here shows four options when you load up a tag file.



- excluded: the tags won't get processed, you'll only see the words
- included: your tagfile tags will be processed:

<noun>ADVANCE	2	0.01%	1	100.00%	0.35	<noun>
<adjective>ADVANCED	1		1	100.00%	0.00	<adjective>
<verb>ADVANCED	1		1	100.00%	0.00	<verb>
<noun>ADVANTAGE	5	0.03%	1	100.00%	0.33	<noun>
<noun>ADVANTAGES	1		1	100.00%	0.00	<noun>
<adjective>ADVERSE	1		1	100.00%	0.00	<adjective>
<noun>ADVERTISING	10	0.06%	1	100.00%	0.55	<noun>
<verb>ADVERTISING	1		1	100.00%	0.00	<verb>
<verb>ADVOCATE	1		1	100.00%	0.00	<verb>
<verb>AFFECT	2	0.01%	1	100.00%	0.35	<verb>
<adjective>AFFORDABLE	1		1	100.00%	0.00	<adjective>
<proper-noun>AFRICA	1		1	100.00%	0.00	<proper-noun>
<adjective>AFRICAN	1		1	100.00%	0.00	<adjective>
<conjunction>AFTER	1		1	100.00%	0.00	<conjunction>

ADVERTISING is both a noun and a verb in this BNC XML text.

- only tags, no words: the words do not get included
- only <tagged> words: words and their tags are included but words which aren't tagged get left out.

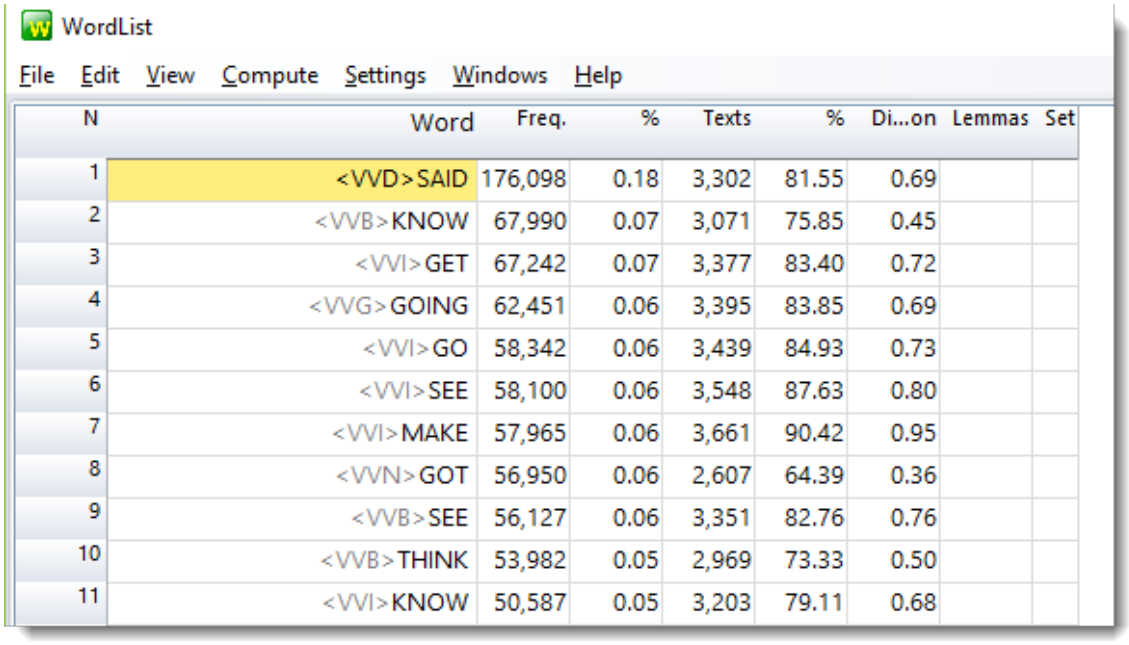
6.7.1 only words with tag-file tags

A setting which lets you restrict a word list to words with neighbouring tags such as **<ADJECTIVE>cheerful**. Only applies if you have loaded up a tag file 172.

In this listing, **OF** is visible but untagged.

<article>THE	918	5.86%	1	100.00%	0.93	<article>
OF	590	3.76%	1	100.00%	0.92	
<preposition>IN	409	2.61%	1	100.00%	0.92	<preposition>
<article>A	383	2.44%	1	100.00%	0.94	<article>
<conjunction>AND	379	2.42%	1	100.00%	0.94	<conjunction>
<verb-be>BE	256	1.63%	1	100.00%	0.84	<verb-be>
<infinitive to>TO	255	1.63%	1	100.00%	0.89	<infinitive to>
<verb-be>IS	235	1.50%	1	100.00%	0.93	<verb-be>
<preposition>FOR	213	1.36%	1	100.00%	0.80	<preposition>
<preposition>TO	173	1.10%	1	100.00%	0.88	<preposition>

This word list, however, only consists of items with a tag matching those in the tag file (of verb tags) used with data derived from the BNC XML edition.



WordList

File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts	%	Di...on	Lemmas	Set
1	<VVD>SAID	176,098	0.18	3,302	81.55	0.69		
2	<VVB>KNOW	67,990	0.07	3,071	75.85	0.45		
3	<VVI>GET	67,242	0.07	3,377	83.40	0.72		
4	<VVG>GOING	62,451	0.06	3,395	83.85	0.69		
5	<VVI>GO	58,342	0.06	3,439	84.93	0.73		
6	<VVI>SEE	58,100	0.06	3,548	87.63	0.80		
7	<VVI>MAKE	57,965	0.06	3,661	90.42	0.95		
8	<VVN>GOT	56,950	0.06	2,607	64.39	0.36		
9	<VVB>SEE	56,127	0.06	3,351	82.76	0.76		
10	<VVB>THINK	53,982	0.05	2,969	73.33	0.50		
11	<VVI>KNOW	50,587	0.05	3,203	79.11	0.68		

Notice that **SEE** and **KNOW** appear twice, distinguishing infinitive and base forms.

6.7.2 tags but no words

Here's an example.

N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
1	<noun>	4,479	28.58%	1	100.00%	0.97		<noun>
2	<adjective>	1,659	10.59%	1	100.00%	0.96		<adjective>
3	<preposition>	1,490	9.51%	1	100.00%	0.95		<preposition>
4	<verb>	1,417	9.04%	1	100.00%	0.97		<verb>
5	<article>	1,386	8.84%	1	100.00%	0.94		<article>
6	<conjunction>	868	5.54%	1	100.00%	0.97		<conjunction>
7		801	5.11%	1	100.00%	0.92		
8	<adverb>	717	4.57%	1	100.00%	0.96		<adverb>
9	<verb-be>	698	4.45%	1	100.00%	0.93		<verb-be>
10	<determiner>	522	3.33%	1	100.00%	0.94		<determiner>
11	<modal>	355	2.27%	1	100.00%	0.85		<modal>
12	<proper-noun>	260	1.66%	1	100.00%	0.86		<proper-noun>
13	<infinitive to>	255	1.63%	1	100.00%	0.89		<infinitive to>
14	<numeral>	237	1.51%	1	100.00%	0.90		<numeral>
15	<initials>	160	1.02%	1	100.00%	0.72		<initials>
16	<pronoun>	156	1.00%	1	100.00%	0.86		<pronoun>
17	<possessive>	111	0.71%	1	100.00%	0.79		<possessive>
18	<verb-have>	76	0.48%	1	100.00%	0.81		<verb-have>
19	<verb-do>	26	0.17%	1	100.00%	0.68		<verb-do>

In line 7, we get 801 cases of words which were not tagged with tag-file mark-up.

6.8 XML text such as BNC XML

What is XML?

XML text has angle-bracketed mark-up which provides additional information. For example the British National Corpus has text which is structured like this:

```
<s n="43">
<w c5="PNP" hw="i" pos="PRON">I </w><w c5="VVB" hw="mean"
pos="VERB">mean</w>
<c c5="PUN">, </c><w c5="AVQ" hw="where" pos="ADV">where </w>
<w c5="VDB" hw="do" pos="VERB">do </w><w c5="NN1-VVG" hw="eating"
pos="SUBST">eating </w>
<w c5="NN2" hw="disorder" pos="SUBST">disorders </w>
<w c5="VVB" hw="come" pos="VERB">come </w><w c5="PRP" hw="from"
pos="PREP">from</w>
<c c5="PUN">?</c>
</s>
```

<s> ... </s> signals a sentence.
<w c5="PNP" hw="i" pos="PRON"> signals that the next word is a pronoun (coded **PNP**), head-word is **"i",**
<w c5="NN2" hw="disorder" pos="SUBST"> signals that the next word is a plural noun belonging to the head-word "**disorder"** and it's a substantive.
c5="NN2" is an ***attribute*** of the **<w** start-tag, **hw="disorder"** is another attribute. There can be many attributes in a start-tag. The **<c** start-tags have only one, but the **<w** start-tags have 3 in this BNC text.

WordSmith's handling of XML

By default, WordSmith simply ignores all the mark-up so a word list will only get the words in black inserted in it, a concordance will only see those words (**I mean, where do eating disorders come from?**).

If you define a tag file with colours you can get views like

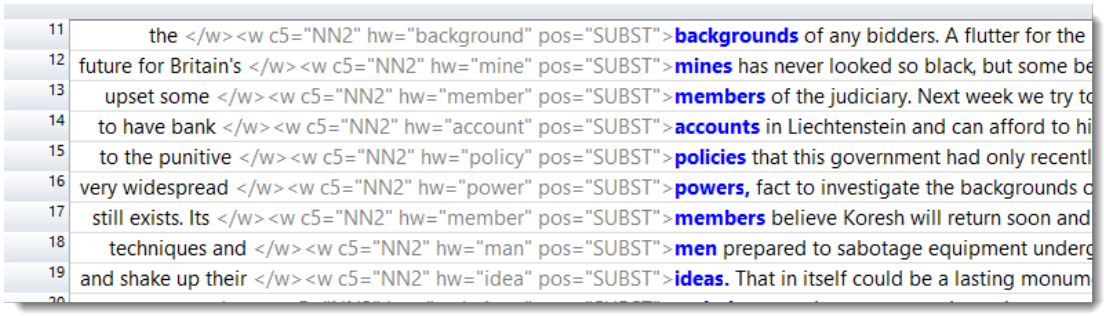
Searching using Attributes

If you want to search for all instances of NN2 forms (plural nouns), you'd need to type

`<w c5="NN2" * *>*`

as your search-word and answer yes to the question as to whether you're concordancing on tags.

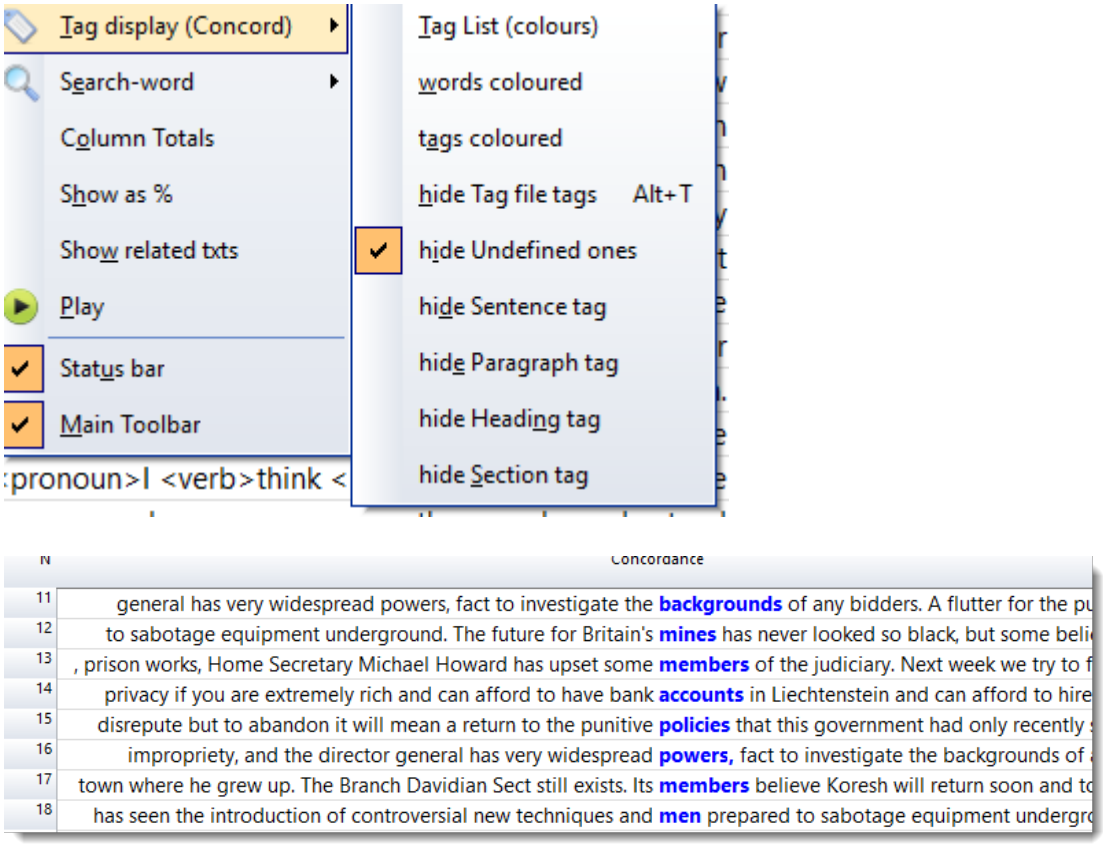
You would get results like this:



Lots of mark-up! See here for a better solution 174 where you define the tags you're interested in first.

Hide the mark-up

If you prefer not to see all that mark-up in grey, choose to hide the undefined mark-up



There is a box in the main tool 257 which can show or hide mark-up, too.

Asterisks in your search-word

In the example above, we search on

<w c5="NN2" * *>*

<w because each start-tag where **NN2** forms are found starts with **<w** and the very first attribute is **c5="NN2"**. Then two asterisks to indicate that we aren't interested in the **hw** or **pos** attributes. Then a closing **>** and another asterisk because the word which follows will be right next to the **>** in our corpus.

For two successive parts of speech,

<w c5="AT0" * *>* <w c5="NN1" * *>*

looks for any article (the/a/an) followed by any singular count noun.

A search on

<w c5="NN?" hw="player" *>*

where we are allowing **NN1** or **NN2** and requiring the **hw** to be **player**,gets results like this:

N	Concordance	S
5	the first half, but catching Bicester cold after a </w><w c5="NN1" hw="player" pos="SUBST">player had	
6	Saturday that it was his shot, parried by , a </w><w c5="NN1" hw="player" pos="SUBST">player with	
7	who's who can, very very good accordion </w><w c5="NN1" hw="player" pos="SUBST">player who	
8	off the line, it was followed up by another </w><w c5="NN1" hw="player" pos="SUBST">player to	
9	number eleven, who looks as quick as any </w><w c5="NN1" hw="player" pos="SUBST">player I've	
10	over the controversial Phil affair. The England </w><w c5="NN1" hw="player" pos="SUBST">player was	
11	I think that speaks erm quite a lot for the erm </w><w c5="NN2" hw="player" pos="SUBST">players	
12	No no, we've got some good experienced </w><w c5="NN2" hw="player" pos="SUBST">players	
13	. Yes, we've got some good experienced </w><w c5="NN2" hw="player" pos="SUBST">players	
14	despite howls of protest from the Fairmile </w><w c5="NN2" hw="player" pos="SUBST">players for	

Another example

Searching Italian .XML containing text like this:

DIALOGICO.xml - Notepad

File Edit Format View Help

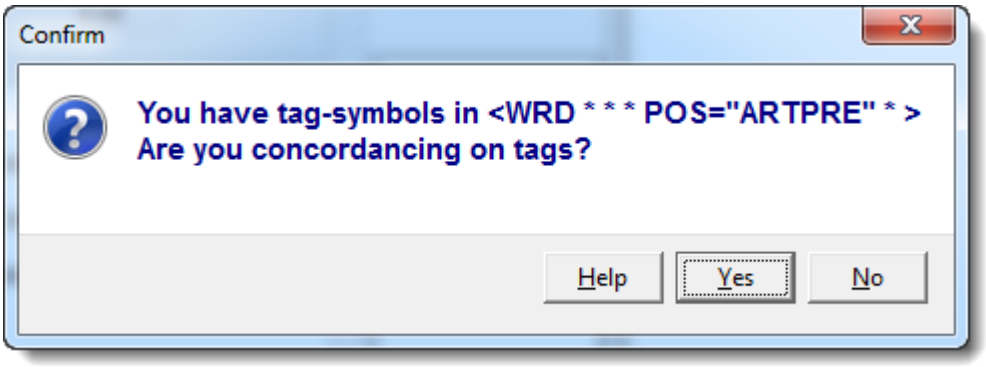
```
<?xml version="1.0" encoding="ISO-8859-1"?>
<RISULTATI>
        <DIALOGO id="DGmt8028" parlanti="2" parole="1254" tempo="523769" >
                <TURNO id="1" >
                        <WRD id="1" start="0" end="1493" POS="NOCAT" lemma="__" >__</WRD>
                        <WRD id="2" start="1493" end="15978" POS="" lemma="" ><inspiration/></WRD>
                        <WRD id="3" start="15978" end="19549" POS="" lemma="" ><sp/></WRD>
                        <WRD id="4" start="19549" end="29593" POS="ADV" lemma="allora" >allora</WRD>
                        <WRD id="5" start="29593" end="39971" POS="" lemma="" ><sp/></WRD>
                        <WRD id="6" start="39971" end="50914" POS="" lemma="" ><ehm/></WRD>
                        <WRD id="7" start="50914" end="60860" POS="VER:fin" lemma="girare" >gira%</WRD>
                        <WRD id="8" start="60860" end="63270" POS="PRE" lemma="a" >%a</WRD>
                        <WRD id="9" start="63270" end="73057" POS="NOUN" lemma="destra" >destra</WRD>
                        <WRD id="10" start="73057" end="75472" POS="ARTPRE" lemma="del" >dei</WRD>
                        <WRD id="11" start="75472" end="88206" POS="NOUN" lemma="limone" >limoni</WRD>
```

and wishing to find all cases of the ARTPRE part of speech, with the search-word specified like this

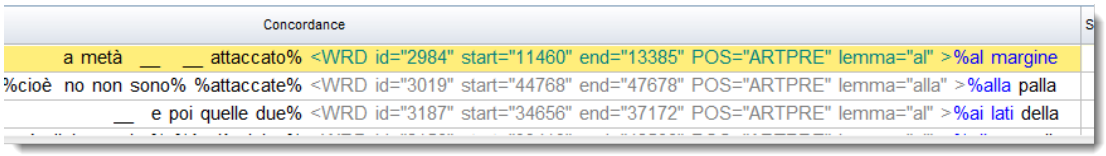
Texts | Search Word | Advanced | Bat

<WRD * * * POS="ARTPRE" * >

and answering yes to this:



we get a considerable concordance with entries like this:



(I have no idea why there are % symbols in the source .XML, by the way.)

See also : Handling the BNC

Concord

7 Concord

Concord is a program which makes a concordance[196] using plain text or web text files.

To use it you will specify a search word or phrase[246], 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[199] them out, then printing[88].

Researchers can use a concordancer, for example when searching through a database of hospital accident records, to see whether *fracture* is associated with *fall*, *grease*, *ladder*. Or to examine historical documents to find all the references to *land ownership*.

Online step-by-step guide showing how

7.1 index

Explanations

What is Concord and what's it for?[195]
Collocation[206]
Collocation Display[210]
Plots[230]
Clusters[203]
Patterns[237]

What to do if it doesn't do what I want... 563

Settings

Choosing texts 52
Collocate horizons 207
Collocate settings 217
Concordance settings 249
Context word 226
Main Controller Concordance Settings 253
Nearest Tag 223
Search word or phrase 246
Tag Concordancing 221
Tagged Texts 157
Text settings 145

Procedures

What you can See and Do 198
Altering the View 257
Blanking Out a Concordance 199
Re-sorting a Concordance 238
Removing Duplicate lines 236
Re-sorting Collocates 219
User-defined categories 43
Editing Concordances 236
Merging Concordances 310
Sound and Video in Concord 258

see also : WordSmith Main Index 2

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 ...
```

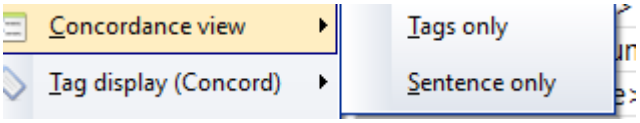
A concordancer searches through a text or a group of texts and then shows the concordance as output. This can be saved, printed, etc.

7.3 concordance view

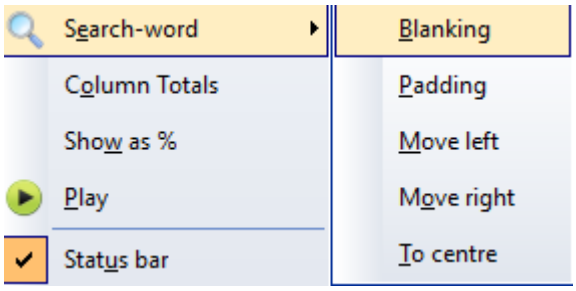
Concordance View

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 60.

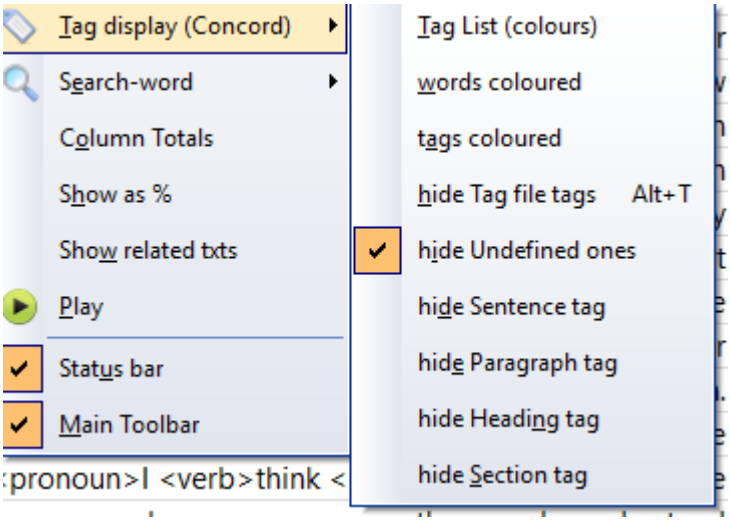
Some of the options are visible here:



sentence only = show the context only up to its left and right sentence boundaries 533. See zero characters 243 if exporting as **txt** or **xlsx**.



Blanking = blank the search-word out e.g. to make a guess-the-word exercise 199
Padding = insert a space to left and right of the search-word so it stands out better



Tag List (colours): a window offering a chance to switch colours on and off 223.
hide undefined tags = hide those not defined in your tag file 172
hide tag file tags = hide all tags including undefined ones
words coloured = colour words
tags coloured = colour the tags
hide the sentence tags = hides e.g. **<s>** and **</s>** tags.
hide the paragraph tags
hide heading tags
hide section tags

Other options are in the Concord what you see settings 257. See also: what can I do with the list 198?
, Saving and printing 242.

7.4 what can I do with the list?

You have a listing showing all the concordance lines in a window. You can scroll up and down and left or right with the mouse or with the cursor keys.

Sort the lines

If you have a lot of lines you should certainly sort them. A concordance is read vertically, not horizontally. You are looking for repeated patterns, such as the presence of lots of the same sorts of words to the right or left of your search-word. Click the bar at the top to start a sort.

The Columns

These show the details for each entry: the entry number, the concordance line, set, tag, word-position (e.g. 1st word in the text is 1), paragraph and sentence position, source text file name, and how far into the file it comes (as a percentage). See below for an explanation of the **purple blobs**. The statistics are computed automatically based on the language settings.

Set

This is where you can classify the entries yourself, using any letter, into user-defined categories 43. 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 246, 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 223.

Other things you may wonder about

Status bar

The status bar 558 panels show

- the number of entries
- whether we're in "Typing" or "Set" mode;
- the current selected cell of the grid.

See also:

Re-sorting 238 your concordance lines
Follow-up 45 searches
User-defined categories 43
Altering the View 257

Blanking out 199 the search-word
Padding the search-word with spaces (use the search-word padding menu item to put a space on each side of the search-word)
Collocation 206 (words in the neighbourhood of the search-word)
Plot 230 (plots where the search-word came in the texts)
Clusters 203 (groups of words in your concordance)
Text segments in Concord 263
Editing the concordance 236
Time-lines 149
Zapping entries 236
Saving and printing 242
Window Management 151

7.4.1 blanking

In a concordance, to blank out the search-words with asterisks, just press the spacebar (or choose *View* | *Blanked out*). Press it again to restore them.

The point of it...

A blanked-out concordance is useful when you want to create an exercise. This one has *give* and *put* mingled:

```
... could not ********** me the time ...
... Rosemary, ********** me another ...
... would not ********** much for that ...
... could not ********** up with him ...
... so you'll ********** him a present ...
... will soon ********** up smoking ...
... he should ********** it over here ...
```

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

See also : Other main Controller settings for Concord 253

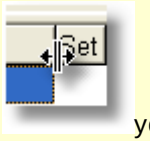
7.4.2 source text context

To see more than a few words on either side of your search-word

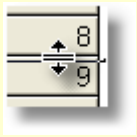
you could press (F8) to "grow" all the rows, or (Ctrl+F8) to shrink them. Or Alt+F8 to grow the current line only (Ctrl+Alt+F8 to shrink it).

Alternatively

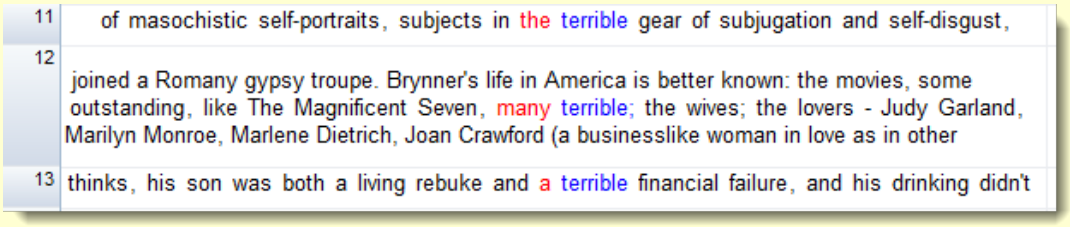
Stretch the display manually

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.

The same applies to each individual row: place the mouse cursor between one row and another in the grey numbered area, and drag.



.



View the whole source text-file

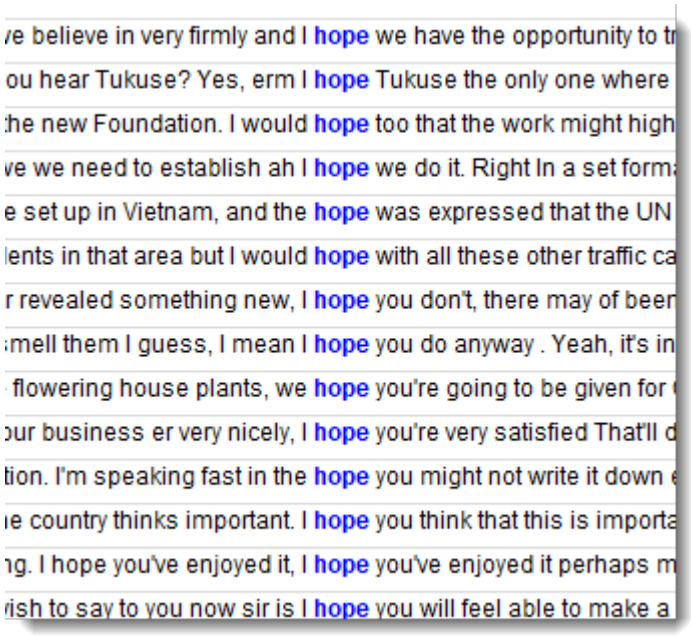
(if it is still on the disk where it was when the concordance was originally created)
Double-click the concordance column, and the source text window will load the file and highlight the search word 134.
Or double-click the file-name column, it will open in Notepad for editing. (If the connection isn't working, re-compute file-lists 242.)

7.5 breakdowns

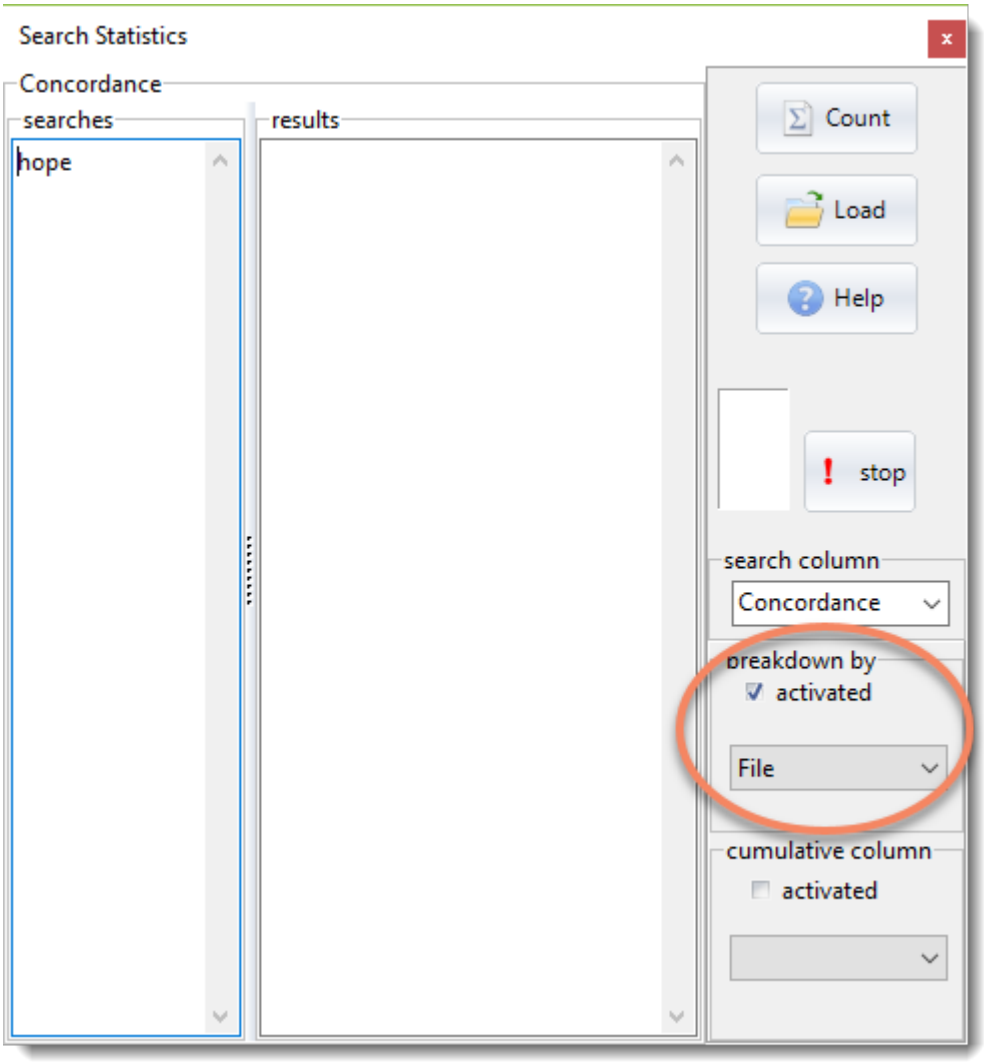
It can be handy to compute a breakdown of the search-word by the different files in your concordance.

Example

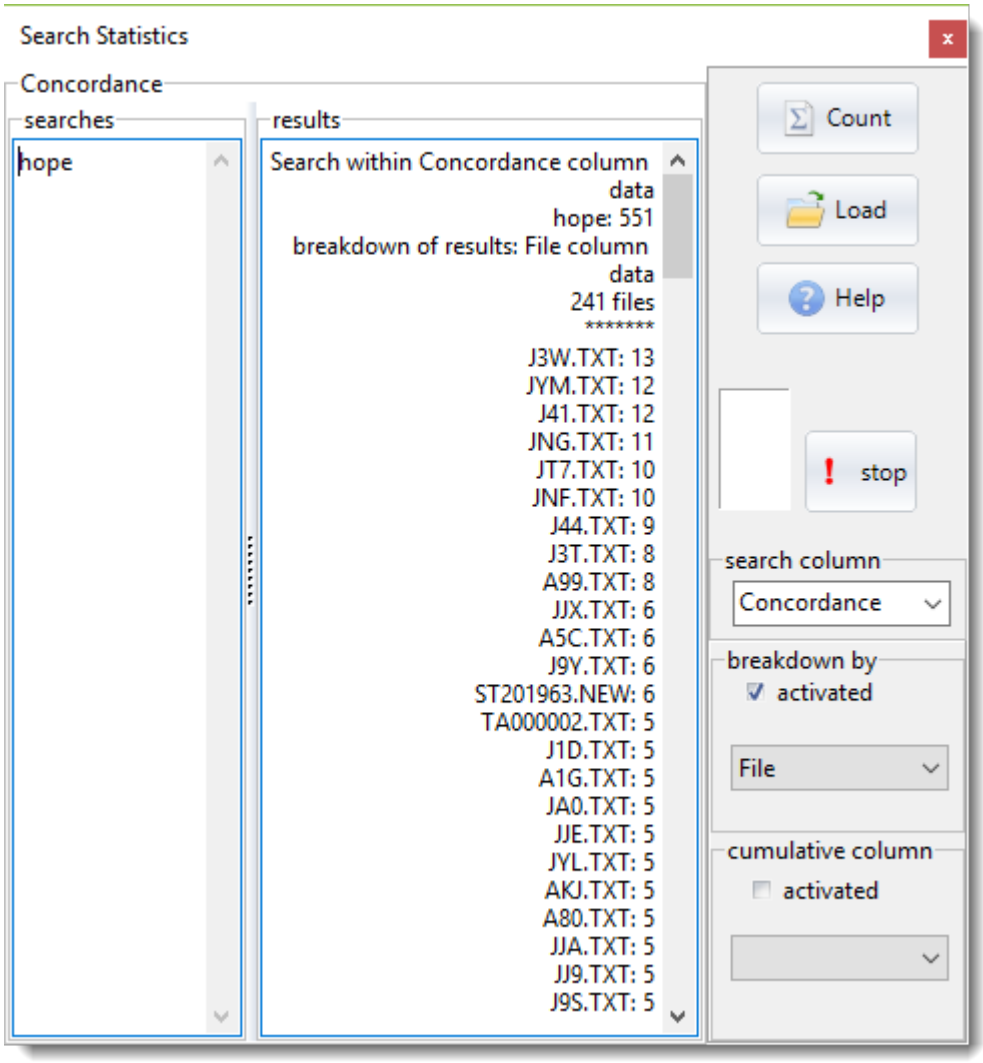
With this concordance, based on lots of text files



choose *Compute* | *Summary Statistics* and you get this.



If you press *Count*, you will get something like this:



showing that there are 551 concordance column entries containing the word *hope*, and 241 text files, *hope* was found 13 times in the concordance entries which referred to text file **J3W.TXT** .

See also: Summary Statistics 138.

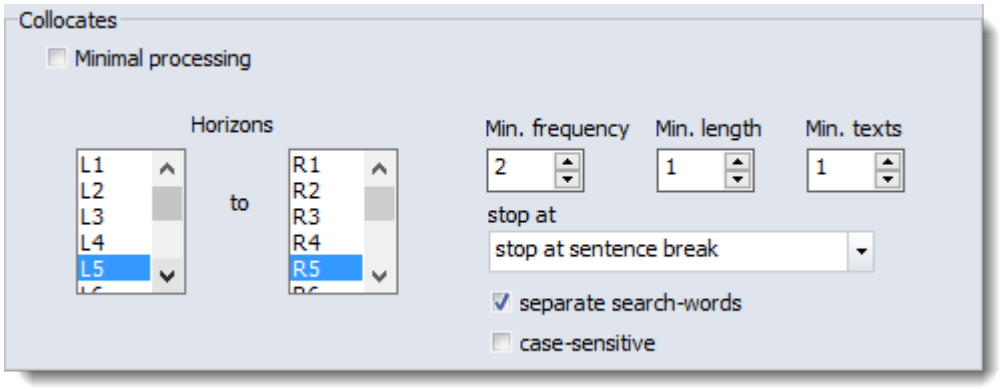
7.6 clusters

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 237.

How it does it…

Clusters are computed automatically if this is not disabled in the main Controller 253 settings for Concord (*Concord Settings*) 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.)

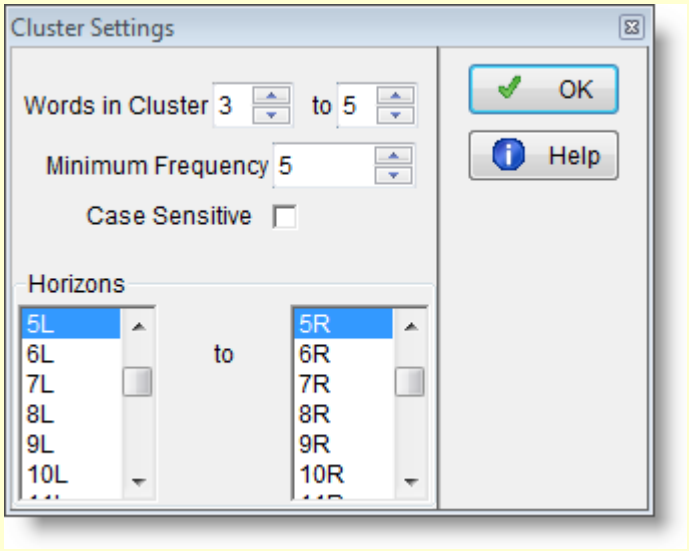
Clusters are sought within these limits: default: 5 words left and right of the search word, but up to 25 left and 25 right allowed. The default is for clusters to be three words in length and you can choose how many of each must be found for the results to be worth displaying (say 3 as a minimum frequency).

Clusters are calculated using the existing concordance lines. That is, any line which has not been deleted or zapped is used for computing clusters.

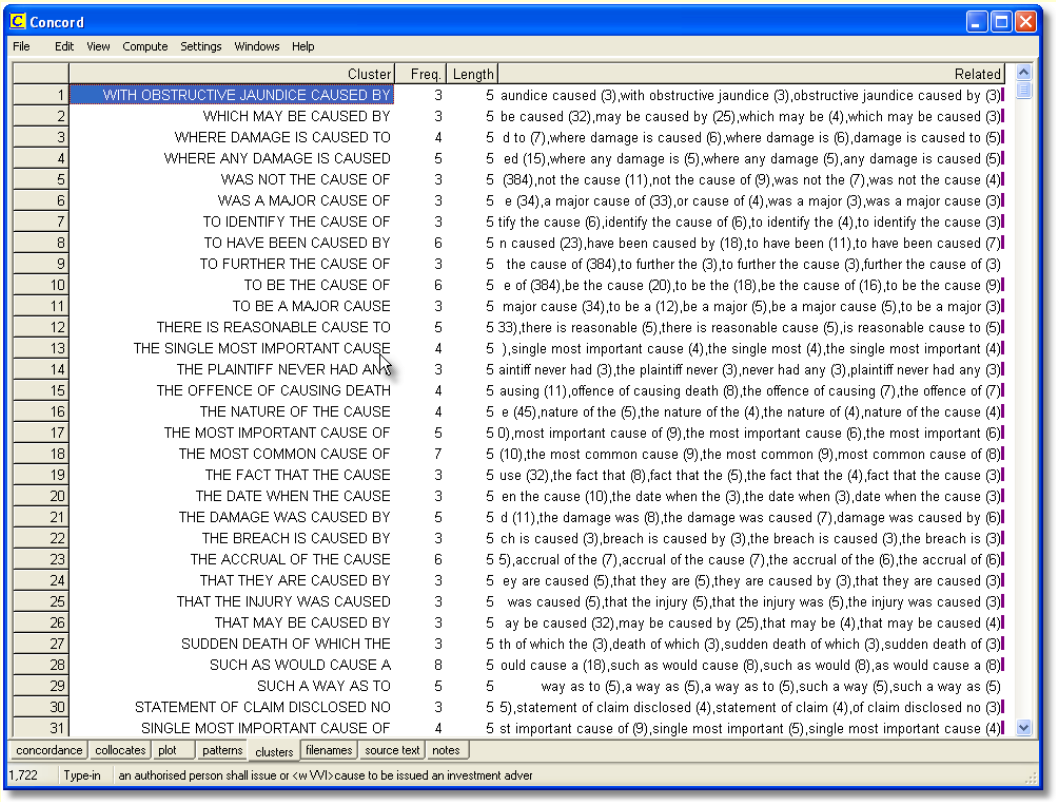
As with WordList index clusters 332, the idea of "stop at sentence breaks 217" (there are other alternatives) is that a cluster which spans across two sentences is not likely to make sense.

Re-computing clusters (../../)

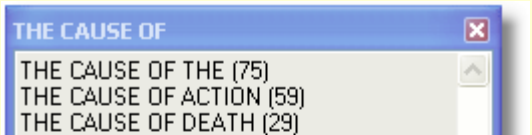
The default clusters computed may not suit, (and you may want to recompute after deleting some lines), so you can also choose *Compute* | *Clusters* ( ) in the Concord menu, so as to choose how many words a cluster should have (cluster size 2 to 4 words recommended), and alter the other settings.



When you press OK, clusters will be computed. In this case we have asked for 3- to 5-word clusters and get results like this:



The clusters have been sorted on the Length column so as to bring the 5-word clusters to the top. At the right there is a set of "Related" clusters, and for most of these lines it is impossible to see all of their entries. To solve this problem, double-click any line in the Related column and another window opens. Here is the window showing what clusters are related to the 3-word cluster, **`the cause of`**, which is the most frequent cluster in this set:



See also: general information on clusters 556, WordList Clusters 332, Word Clouds 152.

7.7 Collocation

7.7.1 what is collocation?

What's a "collocate"?

Collocates are the words which occur in the neighbourhood of your search word. Collocates of *letter* might include *post, stamp, envelope*, etc. However, very common words like *the* will also collocate with *letter*.

and "colligation"?

Linkages between neighbouring words which involve grammatical items are often referred to as *colligation*. That **`rely`** is typically followed by a preposition in English is a colligational fact.

The point of it...

The point of all this is to work out characteristic lexical patterns by finding out which "friends" words typically hang out with. It can be hard to see overall trends in your concordance lines, especially if there are lots of them. By examining collocations in this way you can see common lexical and grammatical patterns of co-occurrence.

Options

You may compute a concordance with or without collocates (*minimal processing* 255): 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 207.
You can re-compute collocates after editing your concordance.
If you want to filter your collocate list, use a match list 82 or stop list 141.
Re-sort 219 a collocate list in a variety of ways.
You can see the strength of relationship 342 between the word and the search-word which the concordance was based on.

Collocates can be viewed 210 after the concordance has been computed.

Technical Note

The literature 524 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 279, which are a pointer to "coherence collocates". **Concord** will supply "neighbourhood collocates". **WordList** and **Concord** allow you also to study relationships between words 342.

See also: collocation display 210, collocation settings 217, collocation relationship 207, relationships between words display 342.

7.7.2 collocate horizons

The collocate horizons represent the number of collocates Concord will find to the left and right of your search word, and the distance used by **KeyWords** in searching out plot-links 292. The default 110 is 5L,5R (5 to left and 5 to right) but you can go up to 25 on either side. You can set whether to set collocation boundaries such as sentence, paragraph breaks 217 too.

To set collocation horizons and other **Concord** settings, choose Concord Main Controller Settings 253, or in the main **WordSmith** Controller 5, choose *Concord Settings*.

See also: Collocate Settings 217

7.7.3 collocation relationship

The point of it...

The idea is to find out *how strongly* each collocate relates to the search-word near which it was found.

Suppose you have made a concordance on the word *love*. You see lots of high frequency collocates, mostly grammar words. These are the top frequency collocates but they are words which collocate with almost any search-word..... What you might want to know is which words are *specially linked* to **love.**

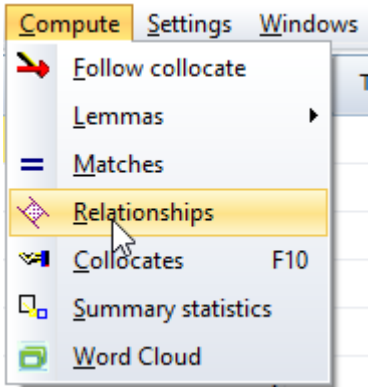
N	Word	Dice	MI	MI3	Z	Log_L	T	Log_R	Con...obA	Con...obB	DeltaPA	DeltaPB
1	LOVE	1.07	12.88	30.16	565.85	5,219.92	20.61	0.00	106.52	106.52	106.52	106.52
2	WITH	0.01	4.66	16.61	7.64	41.18	7.62	-5.47	15.79	0.36	15.16	0.34
3	INCLUSIVENESS	0.01	13.79	14.79	53.22	-0.68	1.41	8.64	0.50	200.00	0.50	199.99
4	SORROWS	0.01	12.21	14.21	30.68	22.45	1.41	7.06	0.50	66.67	0.50	66.65
5	UNREQUITED	0.01	12.21	14.21	30.68	22.45	1.41	7.06	0.50	66.67	0.50	66.65
6	THAT	0.00	2.95	13.67	-1.66	2.97	5.57	-6.56	10.28	0.11	8.93	0.10
7	DISEMBODIED	0.01	11.47	13.47	23.73	19.54	1.41	6.32	0.50	40.00	0.50	39.99
8	TREATISE	0.01	11.47	13.47	23.73	19.54	1.41	6.32	0.50	40.00	0.50	39.99
9	ROMANTIC	0.02	8.69	13.33	14.01	27.74	2.23	2.21	1.25	5.81	1.25	5.80

Here we see them sorted according to their MI3 scores. In this particular corpus we get **Love** and **sorrows**, **love** and **unrequited**. Sad!

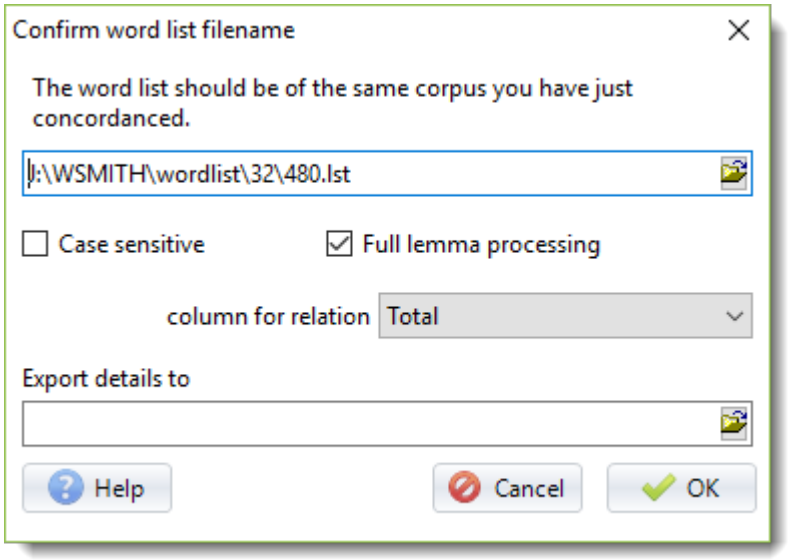
Looking to the right, the DeltaPA 540 score shows a much higher link between **LOVE** and **WITH** than between **WITH** and **LOVE** (DeltaPB 540 score). (If you discuss **love** you may well say who someone is in love **with**, but if you mention **with,** you mostly are not talking about **love**.) And **SORROWS** predicts **LOVE** but **LOVE** doesn't predict **SORROWS**. (Why? Because in this corpus **sorrows** comes only 3 times two of which collocate with **love**, but there are lots of contexts of **love** without **sorrows** (397 out of 399) So maybe it isn't so sad, really.

How to compute it

In the Concord menu, choose *Compute* | *Relationships:*



1. If you haven't done so yet, use WordList to make a word list of the same text files.
2. Now choose the Relationships menu item and Concord will look up each of your collocates in the word list and compute a set of relationship statistics using the information in the reference corpus 555 word list.



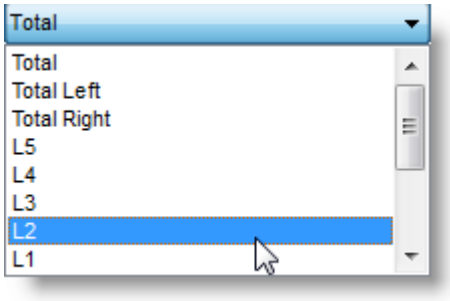
Full lemma processing, case sensitive

These are only relevant if your word list has any lemmatised 322 entries, or it is a case-sensitive word list and you wish processing to respect case-sensitivity. If full lemma processing is checked, each individual joined form will be checked in the relationships computation.

Column for relation

The default is "Total". If you choose Total you're computing the relationship across the current collocation horizons 207 set.

If you prefer to examine the relationship at only one position instead, you may:



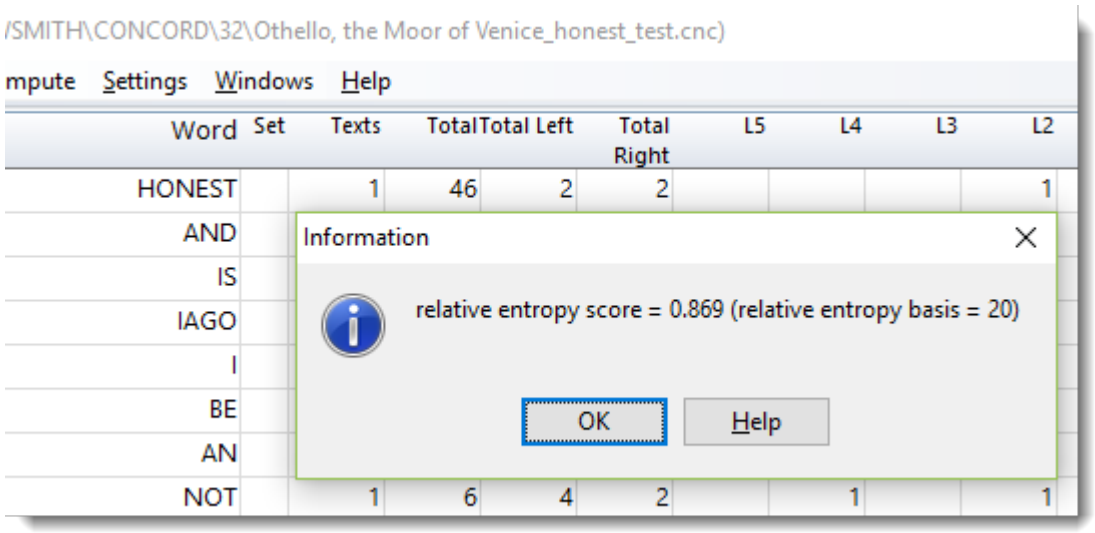
Export Details to

You can optionally choose a file to save details to. This will contain the numbers the statistics are based on and you could take the results into Excel to compute your own statistics.

See also: Collocation 206, Collocate display 210, Mutual Information and other relationship statistics 342, Collocates and relative entropy 209

7.7.4 collocates and relative entropy

Compute the relative entropy of a set of collocates by choosing *Compute* | *Relative entropy* in the menu. This will compute a score and show a result. It will also copy the result into the Notes tab of the concordance.



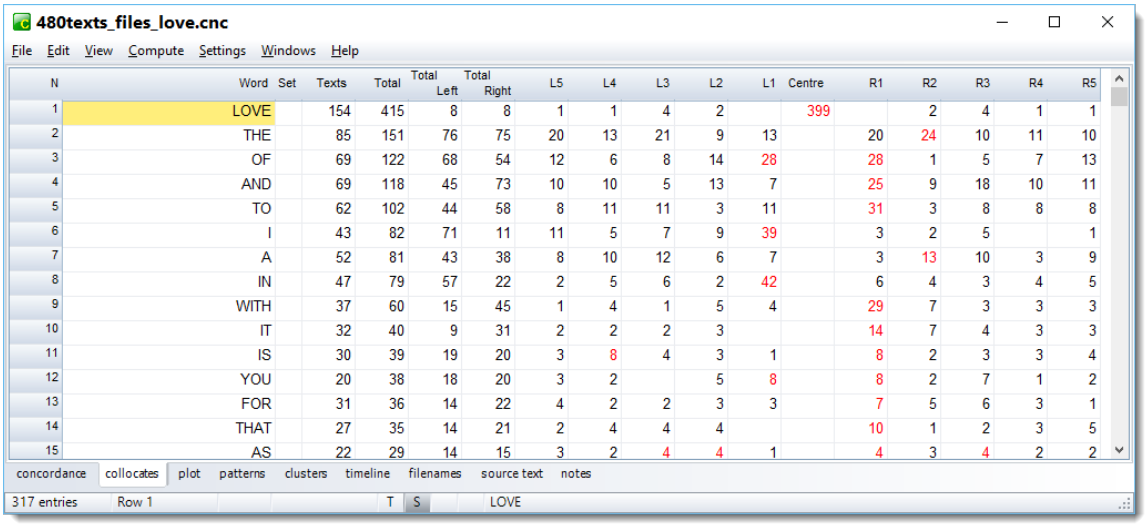
This shows that the set of 20 (of 36) most frequent collocates of **honest** were used to compute a score of 0.869. The scores uses the frequencies shown in the Total column (which spans the range L5 to R5 for these data).

See also: relative entropy 356

7.7.5 collocates display

Display

The collocation display initially shows the collocates in frequency order.



N	Word	Set	Texts	Total	Total Left	Total Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	LOVE		154	415	8	8	1	1	4	2		399		2	4	1	1
2	THE		85	151	76	75	20	13	21	9	13		20	24	10	11	10
3	OF		69	122	68	54	12	6	8	14	28		28	1	5	7	13
4	AND		69	118	45	73	10	10	5	13	7		25	9	18	10	11
5	TO		62	102	44	58	8	11	11	3	11		31	3	8	8	8
6	I		43	82	71	11	11	5	7	9	39		3	2	5		1
7	A		52	81	43	38	8	10	12	6	7		3	13	10	3	9
8	IN		47	79	57	22	2	5	6	2	42		6	4	3	4	5
9	WITH		37	60	15	45	1	4	1	5	4		29	7	3	3	3
10	IT		32	40	9	31	2	2	2	3			14	7	4	3	3
11	IS		30	39	19	20	3	8	4	3	1		8	2	3	3	4
12	YOU		20	38	18	20	3	2		5	8		8	2	7	1	2
13	FOR		31	36	14	22	4	2	2	3	3		7	5	6	3	1
14	THAT		27	35	14	21	2	4	4	4			10	1	2	3	5
15	AS		22	29	14	15	3	2	4	4	1		4	3	4	2	2

You will see the details of how many cases of each collocate come with the search word. The search-word itself will be included, here it's **love**. The word **of** occurs 28 times immediately right of **love**, in L1 position.

The number of words to left and right depends on the collocation horizons 207.
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.

Red numbers

These are the highest numbers within the context horizons. L4 **is** shows up red because 8 occurrences in L4 position is the highest in the range L5 to R5.

Relationships

If you compute relationships you can get to see which collocates are most strongly related to the search-word.

N	Word	Dice	MI	MI3	Z	Log_L	T	Log_R	Set
1	LOVE	1.07	12.93	30.13	568.65	5,094.98	20.37	0.00	
2	I	0.01	4.45	17.16	7.28	39.85	8.64	-6.14	
3	WITH	0.01	4.63	16.45	7.30	37.85	7.43	-5.51	
4	THE	0.00	2.68	17.16	-5.68	36.56	10.37	-8.79	
5	ROMANTIC	0.02	8.76	13.41	14.40	28.26	2.23	2.21	
6	AFFAIRS	0.02	7.90	13.07	11.49	26.82	2.44	1.08	
7	SINGS	0.01	10.83	14.00	23.27	26.10	1.73	5.01	
8	LETTERS	0.02	7.65	12.82	10.44	24.78	2.44	0.83	
9	SORROWS	0.01	12.25	14.25	31.14	22.56	1.41	7.01	
10	UNREQUITED	0.01	12.25	14.25	31.14	22.56	1.41	7.01	
11	FELL	0.02	7.37	12.54	9.36	22.54	2.43	0.55	

New columns appear showing 9 different statistics which might help us find words somehow linked to **love**.

N	Word	Dice	MI	MI3	Z	Log_L	T	Log_R	Con...obA	Con...obB	DeltaPA	DeltaPB
1	LOVE	1.07	12.88	30.16	565.85	5,219.92	20.61	0.00	106.52	106.52	106.52	106.52
2	WITH	0.01	4.66	16.61	7.64	41.18	7.62	-5.47	15.79	0.36	15.16	0.34
3	INCLUSIVENESS	0.01	13.79	14.79	53.22	-0.68	1.41	8.64	0.50	200.00	0.50	199.99
4	SORROWS	0.01	12.21	14.21	30.68	22.45	1.41	7.06	0.50	66.67	0.50	66.65
5	UNREQUITED	0.01	12.21	14.21	30.68	22.45	1.41	7.06	0.50	66.67	0.50	66.65
6	THAT	0.00	2.95	13.67	-1.66	2.97	5.57	-6.56	10.28	0.11	8.93	0.10
7	DISEMBODIED	0.01	11.47	13.47	23.73	19.54	1.41	6.32	0.50	40.00	0.50	39.99
8	TREATISE	0.01	11.47	13.47	23.73	19.54	1.41	6.32	0.50	40.00	0.50	39.99
9	ROMANTIC	0.02	8.69	13.33	14.01	27.74	2.23	2.21	1.25	5.81	1.25	5.80

Same data, different statistic: a rather different impression.

Sorting etc.

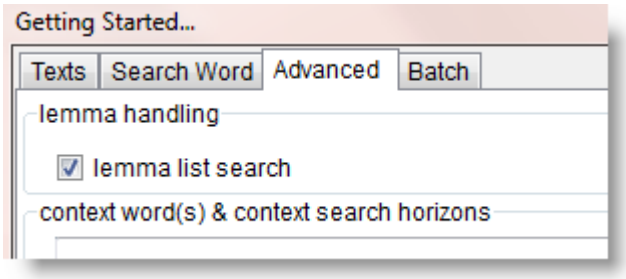
The frequency display can be re-sorted 219 ( ) and you can recalculate the collocates ( ) if you zap 154 entries from the concordance or change the horizons 207.

You can also highlight any given collocate 215 in your concordance display.

See also: Word Clouds 152, Collocation 206, Collocation Relationship 207, Collocates and Lemmas 211, Mutual Information 342

7.7.6 collocates and lemmas

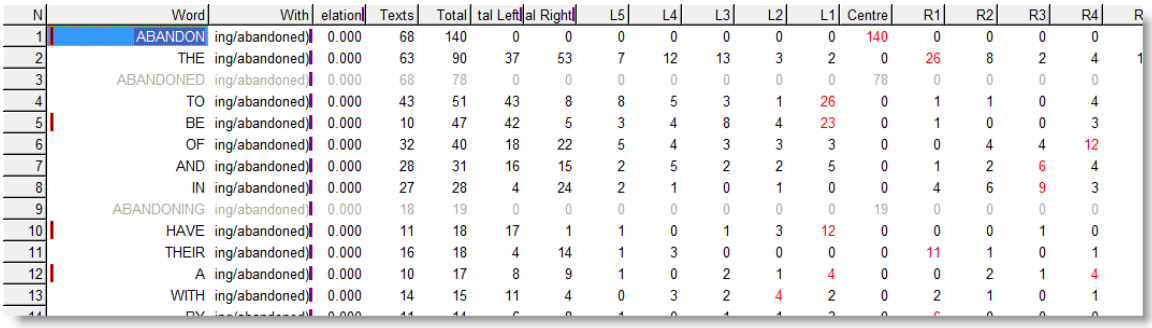
In the following case a lemma list 327 was used and lemma search specified, with a concordance on the word **abandon**:



with these results showing which form of the lemma was used in the Set column.

N	Concordance	Set	Ta
1	to give fish to their fox-farm islands, and abandoned the places. The foxes turned	abandoned	
2	equerry. If that did not feel like being abandoned at birth and pushed out on to	abandoned	
3	and doing another. While they have abandoned the traditional secure	abandoned	
4	broken windows, uprooted plants, abandoned untaxed vehicles, late-night	abandoned	
5	a century - ignored politically by Britain, abandoned by its kith and kin in the	abandoned	
6	that she wasn't coming back. I felt abandoned. My mother had long ago	abandoned	
7	however that might persuade him to abandon his loyalty would be the	abandon	
8	day is nigh, the news filters through of abandoned rafts. Cuban rescue pilots in	abandoned	
9	the 20-club super-league myopically abandoned after the 1990 World Cup.	abandoned	
10	get decent procurement. The directions abandon many of Whitehall's worst	abandon	
11	Owen Chadwick, people all over Europe abandoned their homes in terror and ran	abandoned	
12	around a tree-lined main street and an abandoned 16th-century palace set on	abandoned	
13	world needs is to reinvent the bargain it abandoned. Instead of floating currencies	abandoned	

In the collocate window below, the red line in row 1 indicates that the 140 cases of **ABANDON** include other forms such as 78 cases of **ABANDONED** and 19 of **ABANDONING** (greyed out below).



N	Word	With	elation	Texts	Total	tal Left	al Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R
1	ABANDON	ing/abandoned	0.000	68	140	0	0	0	0	0	0	0	140	0	0	0	0	
2	THE	ing/abandoned	0.000	63	90	37	53	7	12	13	3	2	0	26	8	2	4	1
3	ABANDONED	ing/abandoned	0.000	68	78	0	0	0	0	0	0	0	78	0	0	0	0	
4	TO	ing/abandoned	0.000	43	51	43	8	8	5	3	1	26	0	1	1	0	4	
5	BE	ing/abandoned	0.000	10	47	42	5	3	4	8	4	23	0	1	0	0	3	
6	OF	ing/abandoned	0.000	32	40	18	22	5	4	3	3	3	0	0	4	4	12	
7	AND	ing/abandoned	0.000	28	31	16	15	2	5	2	2	5	0	1	2	6	4	
8	IN	ing/abandoned	0.000	27	28	4	24	2	1	0	1	0	0	4	6	9	3	
9	ABANDONING	ing/abandoned	0.000	18	19	0	0	0	0	0	0	0	19	0	0	0	0	
10	HAVE	ing/abandoned	0.000	11	18	17	1	1	0	1	3	12	0	0	0	1	0	
11	THEIR	ing/abandoned	0.000	16	18	4	14	1	3	0	0	0	0	11	1	0	1	
12	A	ing/abandoned	0.000	10	17	8	9	1	0	2	1	4	0	0	2	1	4	
13	WITH	ing/abandoned	0.000	14	15	11	4	0	3	2	4	2	0	2	1	0	1	

The red mark by **BE** (row 5) shows that this row gives collocation numbers covering all forms of BE such as WAS, WERE etc. Similarly, **HAVE** and **A** are lemmatised in this screenshot.

Thus, for your search-word and its variants you can see detailed frequencies, but its collocates, though they do get lemmatised, do not show you the variant forms or any specific frequencies.

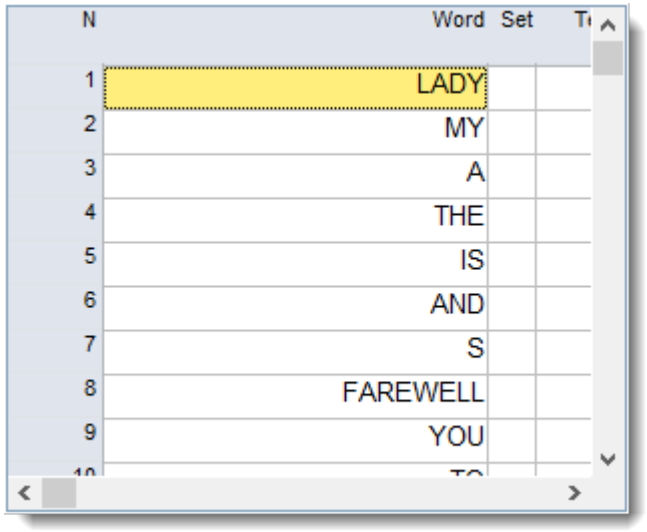
7.7.7 collocate follow

The point of it

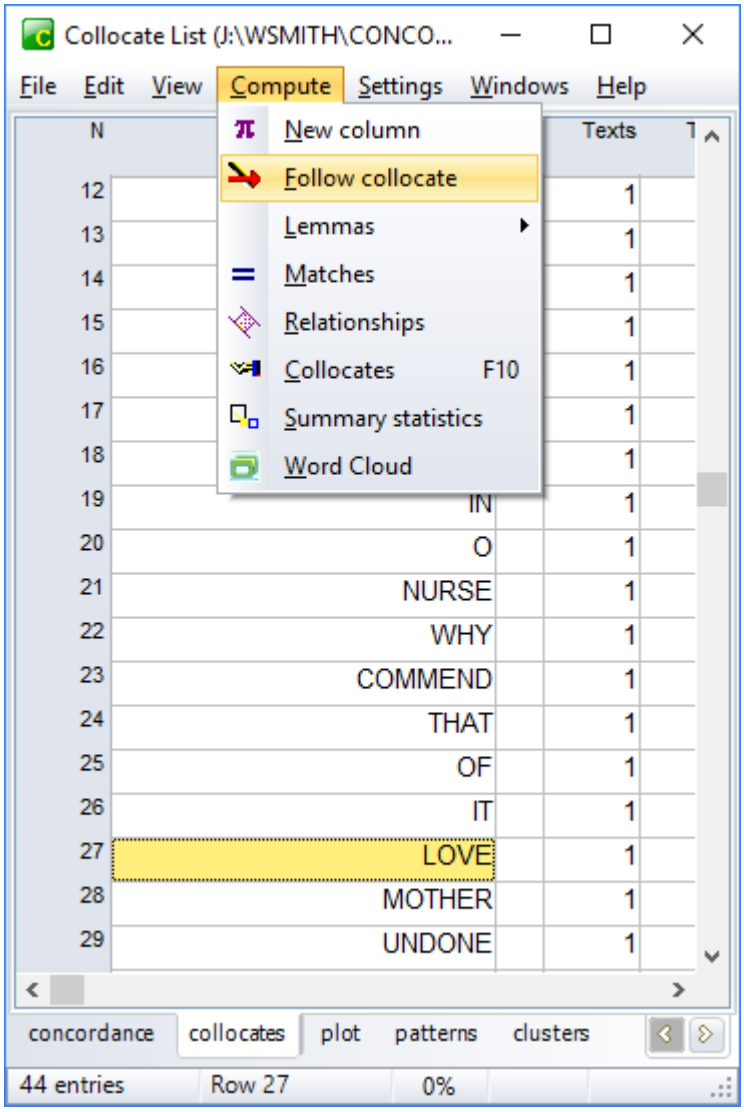
The idea (from Paul Raper) is to be able to follow up a collocate by requesting a new concordance based on it, in the same text files as selected for the collocate. This aids exploration of related words.

How to do it

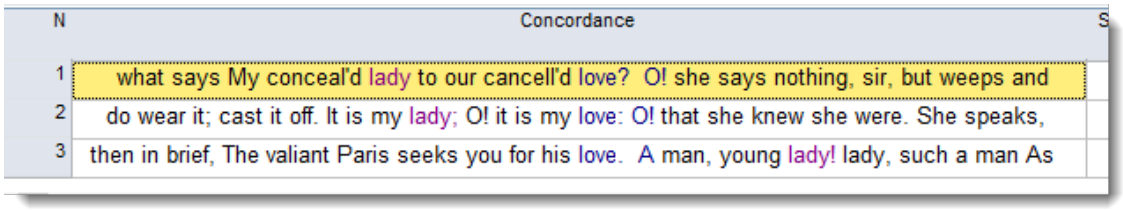
Here is an example, where there is a collocate list relating to **LADY** (from *Romeo and Juliet*).



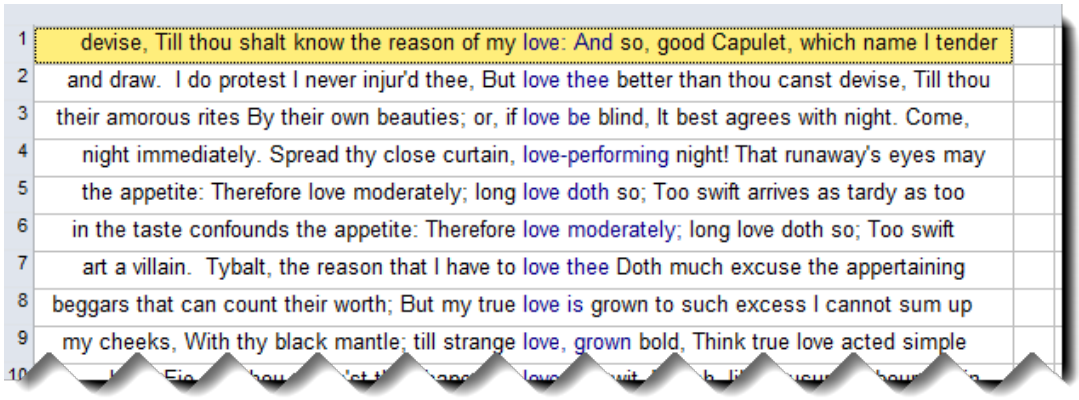
Select a word of interest, such as **LOVE,** and select *Follow collocate* in the *Compute* menu.



You will now be asked whether you want a concordance of the word selected (here **LOVE**) using the same text file(s). If you agree you'll also be asked whether you want to restrict the search so that it finds **LOVE** only in contexts similar to your previously concordance (**LADY** nearby). If you do want it restricted you get results like



and if not, a search on LOVE in the same text file(s):



The search is carried out on the most recently selected text files (selected using the file-choose window 52 or by reading in a saved concordance).

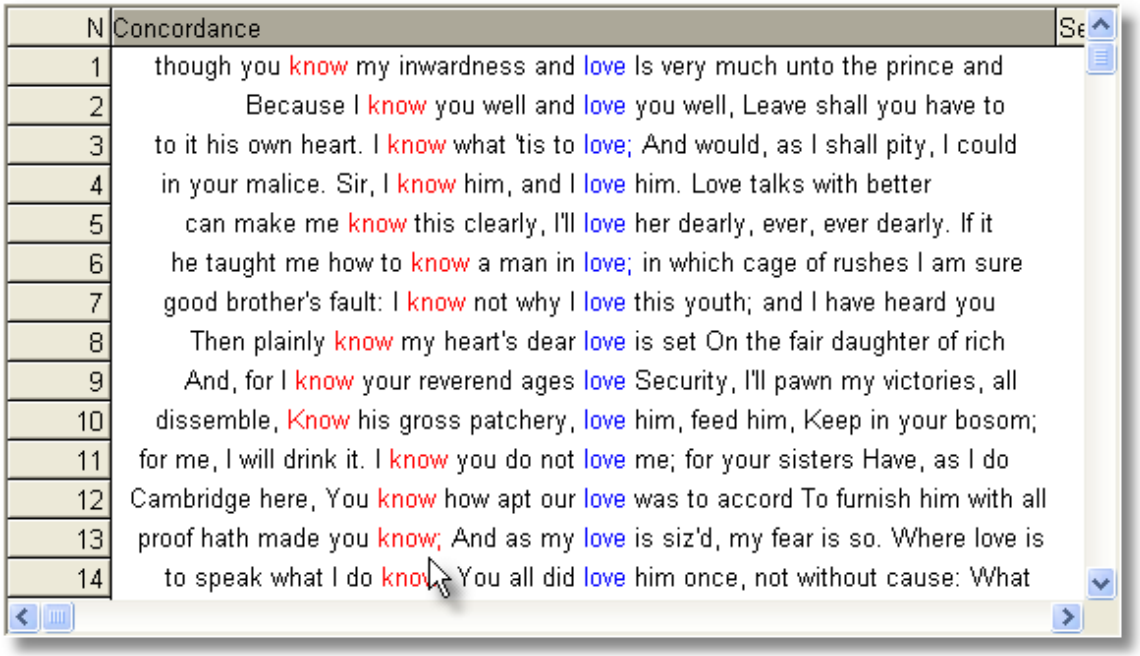
7.7.8 collocate highlighting in concordance

The point of it...

The idea is to be able to see a selected collocate highlighted in the concordance. In this example, the texts were Shakespeare plays and search word was **love**. One of the collocates is **know**, occurring a total of 50 times, with the most frequent at position 4 words to the left of **love**.

N	Word	Total	tal Left	al Right	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
53	KNOW	50	33	17	9	14	2	7	1	0	2	4	1	6	4
54	WHICH	49	14	35	3	0	3	6	2	0	17	2	5	5	6
55	THEY	49	26	23	1	3	7	0	15	0	7	2	1	6	7
56	TRUE	49	35	14	2	5	5	5	18	0	0	2	1	6	5
57	WOULD	46	15	31	4	5	1	1	4	0	2	12	7	8	2

Double-clicking **14** in the L4 column to the right of **know**, we see this in the concordance:



N	Concordance	Se
1	though you know my inwardness and love Is very much unto the prince and	
2	Because I know you well and love you well, Leave shall you have to	
3	to it his own heart. I know what 'tis to love; And would, as I shall pity, I could	
4	in your malice. Sir, I know him, and I love him. Love talks with better	
5	can make me know this clearly, I'll love her dearly, ever, ever dearly. If it	
6	he taught me how to know a man in love; in which cage of rushes I am sure	
7	good brother's fault: I know not why I love this youth; and I have heard you	
8	Then plainly know my heart's dear love is set On the fair daughter of rich	
9	And, for I know your reverend ages love Security, I'll pawn my victories, all	
10	dissemble, Know his gross patchery, love him, feed him, Keep in your bosom;	
11	for me, I will drink it. I know you do not love me; for your sisters Have, as I do	
12	Cambridge here, You know how apt our love was to accord To furnish him with all	
13	proof hath made you know; And as my love is siz'd, my fear is so. Where love is	
14	to speak what I do know. You all did love him once, not without cause: What	

We have brought to the top of the concordance those lines which contain **know** in position L4.

How to do it

In a collocates window or a patterns window, simply double-click the item you wish to highlight. Or select it and choose *View* | *Highlight selected collocate*.

In the collocates window, if you click

	what you get
the Word column or the Total column	all instances of the word
Total Left	those to the left (33 in the case of **know** above)
Total Right	those to the right (17)
otherwise	those in that column only

To get rid

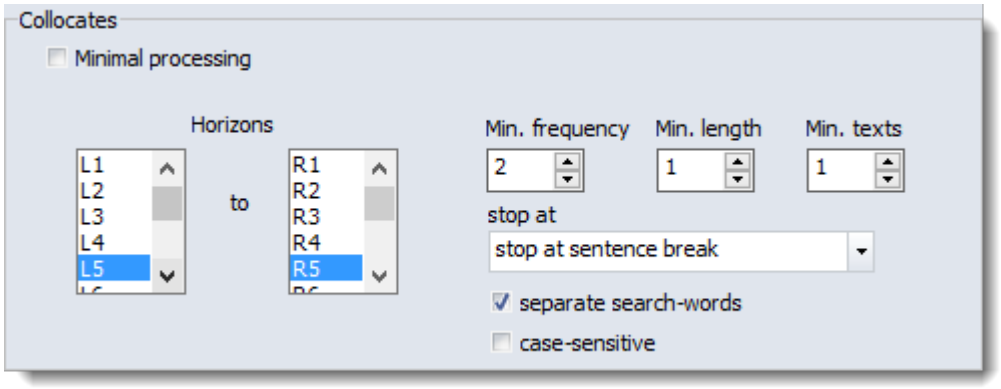
Re-sort 238 in a different way or choose the menu item *View* | *Refresh*.

7.7.9 follow-up concordances

You may wish to chase up some combinations of search-word and collocate.

7.7.10 collocate settings

To set collocation horizons and other **Concord** settings, in the main **WordSmith** Controller 5 menu at the top, choose *Concord Settings*.



Collocates are computed case-insensitively (so **my** in the concordance line will be treated like **My**) unless you check the case-sensitive box.
If you don't want certain collocates such as **THE** to be included, use a stop-list 141.
Numbers will be summarised as # or not according to your language settings 79.
You can lemmatise (join related forms like **SPEAK -> SPEAKS, SPOKE, SPOKEN**) using a lemma list file 327.

Minimum Specifications

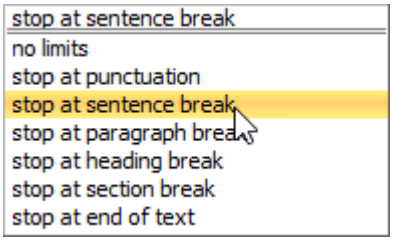
The minimum length is 1, and minimum frequency is 1 (default is 10). You can specify here how frequently it must have appeared in the neighbourhood of the Search Word. Words which only come once or twice are less likely to be informative. So specifying 5 will only show a collocate which comes 5 or more 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. And a minimal number of texts so as to avoid words which collocate only in one or two texts.

Horizons

Here you specify how many words to left and right of the Search Word are to be included in the collocation search: the size of the "neighbourhood" referred to above. The maximum is 25 left and 25 right. Results will later show you these in separate columns so you can examine exactly how many times a given collocate cropped up say 3 words to the left of your Search Word.
The most frequent will be signalled in the *most frequent collocate colour* 114 (default=**red**).

Breaks

These are



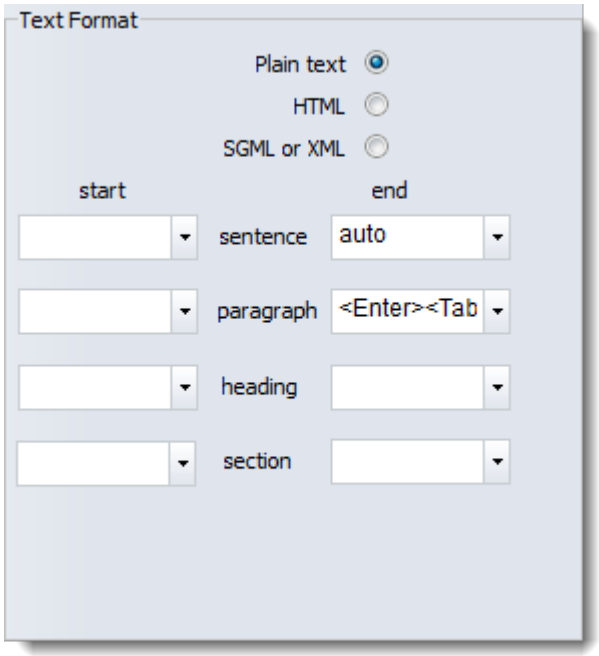
which you will see in the bottom right corner of the screen visible in the Controller Concord Settings 253.
When the collocates are computed, if the setting is to stop at sentence breaks, collocates will be counted within the above horizons but taking sentence breaks into account.

For example, if a concordance line contains

source, per pointing integration times, respectively. However, when we compared these two maps

and the search-word is **however**,
only
when we compared these two
will be used for collocates because there is a sentence break to the left of the search word. If the setting is "stop at punctuation", then nothing will come into the collocate list for that line (because there is a more major break than punctuation to the left of it, and no word to the right of the search-word before a punctuation symbol.

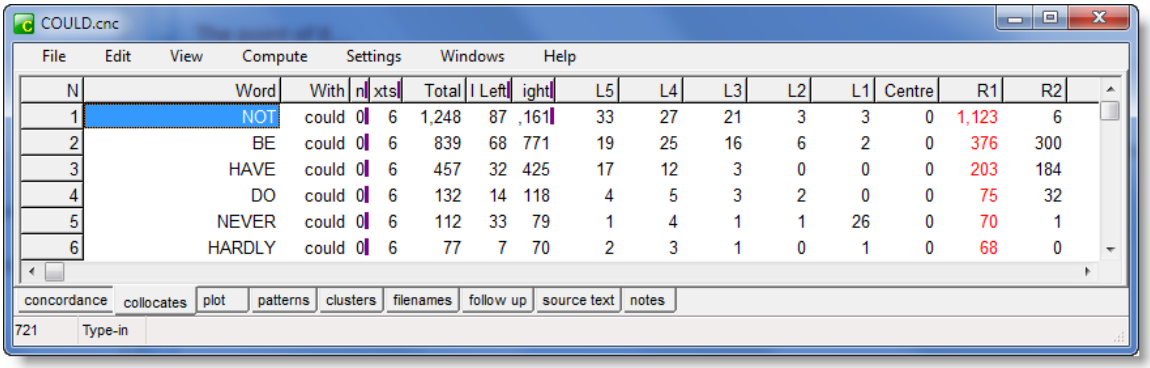
stop at end of text: end of text is by default assumed to be the end of the text file. *stop at heading* or *section*: this works by recognising ends of heading or section which you can specify in the text format box (language settings):



7.7.11 re-sorting: collocates

The point of it…

is to home in, for example, on the ones in L1 or R1 position. To find sub-patterns of collocation, so as to more fully understand the company your search-word keeps.



N	Word	With	nl	xts	Total	l Left	ight	L5	L4	L3	L2	L1	Centre	R1	R2
1	NOT	could	0	6	1,248	87	,161	33	27	21	3	3	0	1,123	6
2	BE	could	0	6	839	68	771	19	25	16	6	2	0	376	300
3	HAVE	could	0	6	457	32	425	17	12	3	0	0	0	203	184
4	DO	could	0	6	132	14	118	4	5	3	2	0	0	75	32
5	NEVER	could	0	6	112	33	79	1	4	1	1	26	0	70	1
6	HARDLY	could	0	6	77	7	70	2	3	1	0	1	0	68	0

Here the collocates of **COULD** in some Jane Austen texts show how negatives crop up a lot in R1 position.

How to do it... just press the header

The frequency-ordered collocation display can be re-sorted to reveal the frequencies sorted by their total frequencies overall (the default), by the left or right frequency total, or by any individual frequency position. Just press the header of a column to sort it. Press again to toggle the sort between ascending and descending.

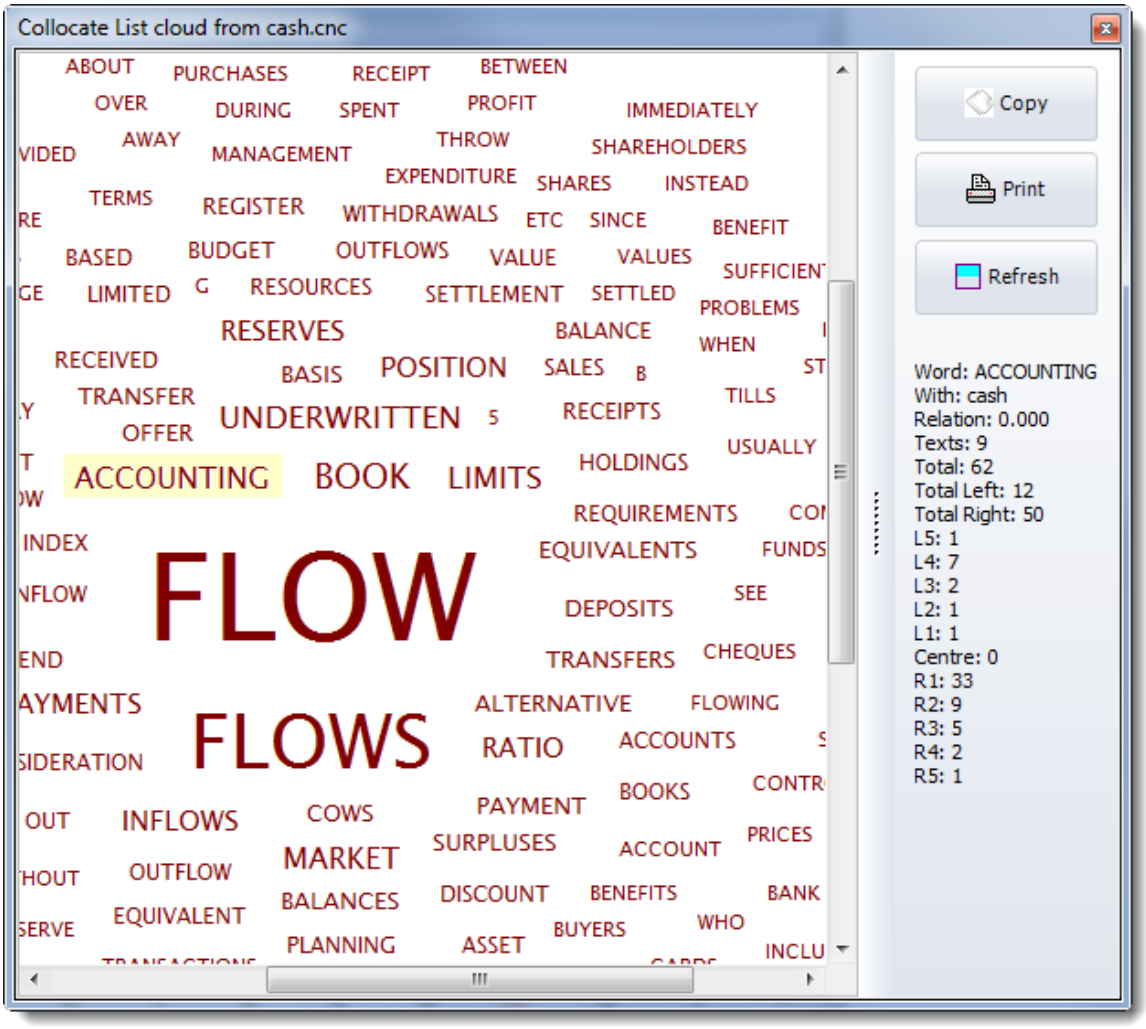
You can also get the concordance lines sorted so as to highlight specific collocates 215, as in the case of the 70 cases of **NEVER** in R1 position in the screenshot.

Word Clouds

You can also get a word cloud 152 of your sorted column. In the screenshot below, a concordance on **cash** generated these R1 collocates (with most function words eliminated using a stoplist 141):

N	Word	R1
1	FLOW	303
2	FLOWS	166
3	BOOK	41
4	UNDERWRITTEN	37
5	ACCOUNTING	33
6	MARKET	31
7	LIMITS	27
8	RATIO	24
9	POSITION	22
10	PAYMENTS	20
11	RESERVES	20
12	INFLOWS	19
13	ALTERNATIVE	18
14	COWS	18
15	BALANCES	17
16	SURPLUSES	15
17	BASIS	14
18	PAYMENT	13
19	TRANSFER	13
20	DEPOSITS	12
21	EQUIVALENTS	12
22	OFFER	12
23	TRANSFERS	12
24	RESOURCES	11
25	INDEX	10
26	RECEIVED	10

and these data fed straight into a word cloud.



In the word cloud, the mouse hovered over the word **accounting** so the details of that word are shown to the right.

See also: Collocation 206, Collocation Display 210, Collation Horizons 207, Word Clouds 152, Patterns 237.

7.8 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 172.

Examples

It doesn't matter whether you are using a tag file 172 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.)

Are you concordancing on tags?

If you are asked this and your search-word or phrase includes tags, answer "Yes" to this question. If not, your search word will get " " inserted around each < or > symbol in it, as explained under Search Word Syntax 246.

Case Sensitivity

Tags are only case sensitive if your search-word or phrase is. Search words aren't (by default). So in example 1, you will retrieve *table* and *Table* and *TABLE* if used as nouns (but nothing at all if no tags are in your source texts).

Hide Tags?

After you have generated a concordance you may wish to hide 224 the mark-up. See the *View* menu for this.

See also: Overview of Tags 157, Handling Tags 158, Showing Nearest Tags in Concord 223, Search word or phrase 246, Types of Tag 159, Viewing the Tags 401, Using Tags as Text Selectors 167

7.8.1 nearest tag

Concord allows you to see the nearest tag, if you have specified a tag file 172, which teaches WordSmith Tools what your preferred tags are. Then, with a concordance on screen, you'll see the tag in one of the columns of the concordance window.

The point of it...

The advantage is that you can see how your concordance search-word relates to marked-up text. For example, if you've tagged all the speech by Robert as **[Rob]** and Mary as **[Mary]**, you can quickly see in any concordance involving conversation between Mary, Robert and others, which ones came from each of them.
Alternatively, you might mark up your text as **<Introduction>, <Body>** and **<Conclusion>**:
Nearest Tag will show each line like this:

```
 1 ... could not give me the time ...   <Introduction>
 2 ... Rosemary, give me another ...           <Body>
 3 ... wanted to give her the help ...         <Body>
 4 ... would not give much for that ...   <Conclusion>
```

To mark up text like this, make up a tag file 172 with your sections and label them as sections, as in these examples:

```
<ABSTRACT> /description "section"
</ABSTRACT>
<INTRODUCTION> /description "section"
</INTRODUCTION>
<SECTION 1> /description "section"
</SECTION 1>
```

or, if you want to identify the speech of all characters in a play, and have a list of the characters, and they are marked up appropriately in the text file, something like this:

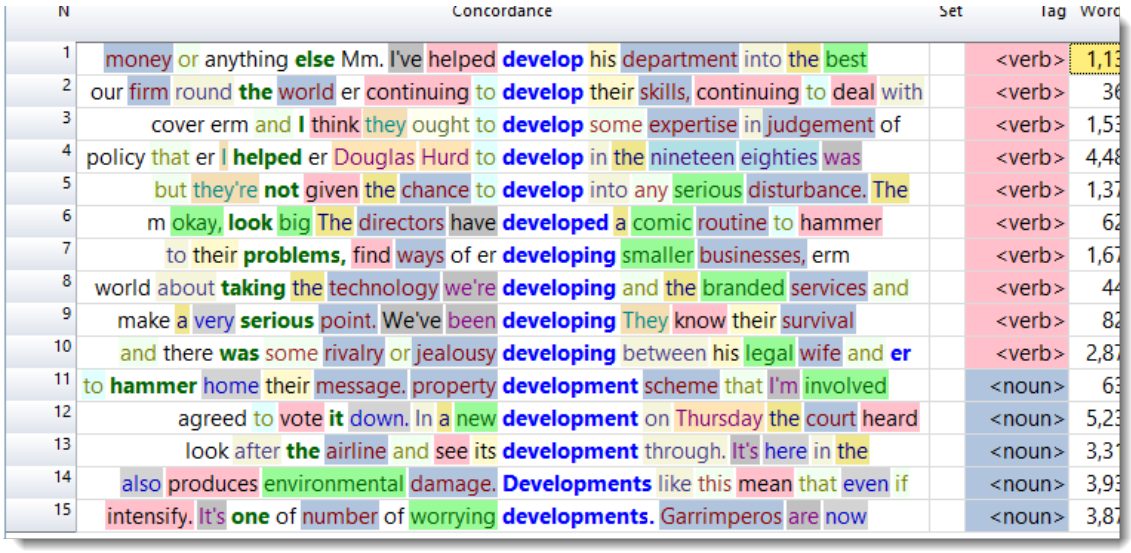
```
<Romeo> /description "section"
</Romeo>
<Mercutio> /description "section"
</Mercutio>
<Benvolio> /description "section"
</Benvolio>
```

In cases using "section", Nearest Tag will find the section, however remote in the text file it may be. Without the keyword "section", Nearest Tag shows only the current context within the span of text saved 255 with each concordance line.

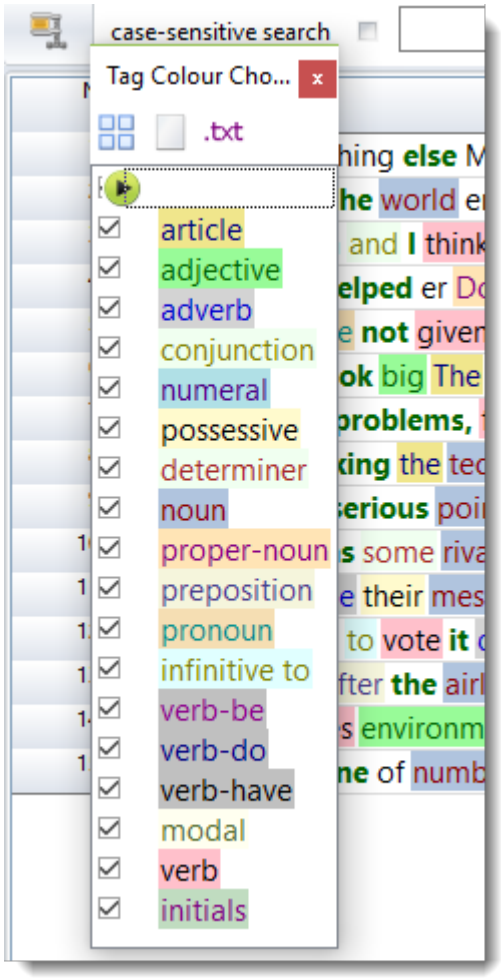
You can sort 238 on the nearest tags. In the shot below, a concordance of nouns has been computed using BNC XML text.

N	Concordance	Set	Tag	Word #
1	I<verb-have>'ve <verb>helped <verb>develop <possessive>his <noun>		<verb>	1,139
2	continuing <infinitive to>to <verb>develop <possessive>their <noun>skills,		<verb>	364
3	<modal>ought <infinitive to>to <verb>develop <determiner>some <noun>		<verb>	1,538
4	Hurd <infinitive to>to <verb>develop <preposition>in <article>the		<verb>	4,486
5	<noun>chance <infinitive to>to <verb>develop <preposition>into <determiner>		<verb>	1,375
6	directors <verb-have>have <verb>developed <article>a <adjective>comic		<verb>	621
7	, <verb>find <noun>ways of er <verb>developing <adjective>smaller <noun>		<verb>	1,670
8	<pronoun>we<verb-be>'re <verb>developing <conjunction>and <article>		<verb>	449
9	<verb-have>'ve <verb-be>been <verb>developing <pronoun>They <verb>know		<verb>	820
10	<conjunction>or <noun>jealousy <verb>developing <preposition>between		<verb>	2,873
11	message. <noun>property <noun>development <noun>scheme		<noun>	636
12	In <article>a <adjective>new <noun>development <preposition>on		<noun>	5,238
13	and <verb>see <possessive>its <noun>development <preposition>through.		<noun>	3,318
14	environmental <noun>damage. <noun>Developments <preposition>like		<noun>	3,939
15	number of <adjective>worrying <noun>developments. <noun>Garrimperos		<noun>	3,877

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



N	Concordance	Set	Tag	Word #
1	money or anything **else** Mm. I've helped **develop** his department into the best		<verb>	1,13
2	our firm round **the** world er continuing to **develop** their skills, continuing to deal with		<verb>	36
3	cover erm and **I** think they ought to **develop** some expertise in judgement of		<verb>	1,53
4	policy that er **I** **helped** er Douglas Hurd to **develop** in the nineteen eighties was		<verb>	4,48
5	but they're **not** given the chance to **develop** into any serious disturbance. The		<verb>	1,37
6	m okay, **look** big The directors have **developed** a comic routine to hammer		<verb>	62
7	to their **problems,** find ways of er **developing** smaller businesses, erm		<verb>	1,67
8	world about **taking** the technology we're **developing** and the branded services and		<verb>	44
9	make a very **serious** point. We've been **developing** They know their survival		<verb>	82
10	and there **was** some rivalry or jealousy **developing** between his legal wife and **er**		<verb>	2,87
11	to **hammer** home their message. property **development** scheme that I'm involved		<noun>	63
12	agreed to vote **it** down. In a new **development** on Thursday the court heard		<noun>	5,23
13	look after **the** airline and see its **development** through. It's here in the		<noun>	3,31
14	also produces environmental damage. **Developments** like this mean that even if		<noun>	3,93
15	intensify. It's **one** of number of worrying **developments.** Garrimperos are now		<noun>	3,87



You can check or uncheck the colours in the list:

Here only adjectives adverbs and verbs have been chosen to be coloured.

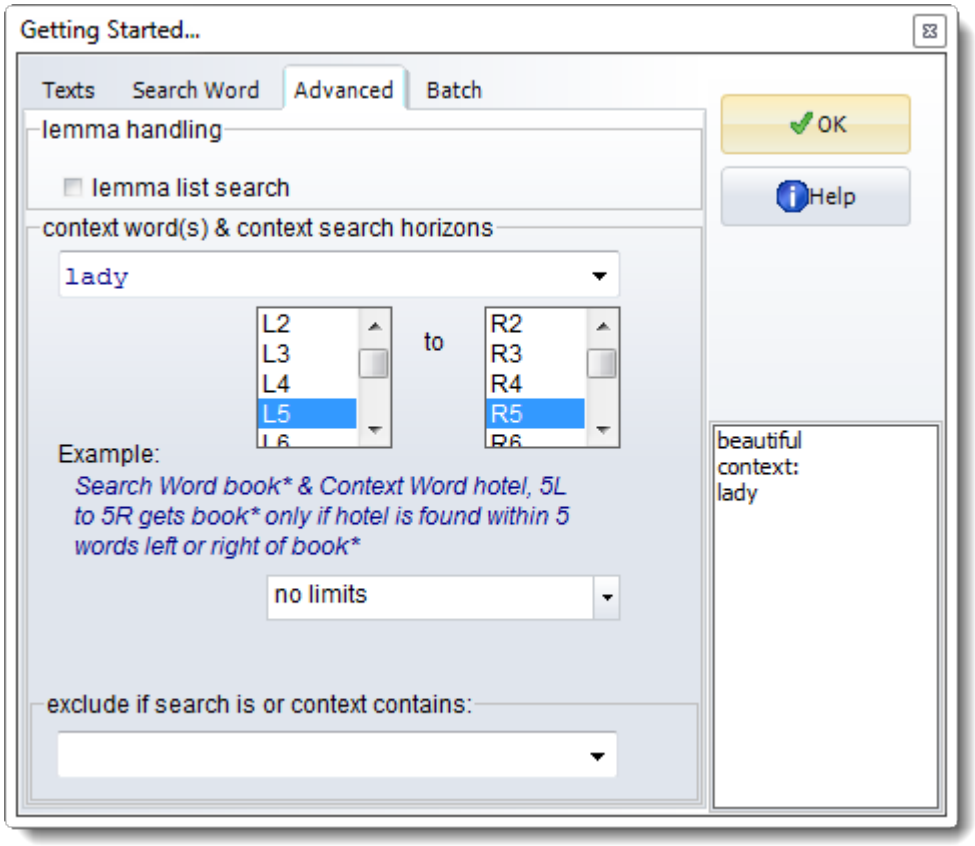
See also: Guide to handling the BNC, Overview of Tags 157, Handling Tags 158, Making a Tag File 172, Types of Tag 159, Viewing the Tags 401, Using Tags as Text Selectors 167

7.9 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 110 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.

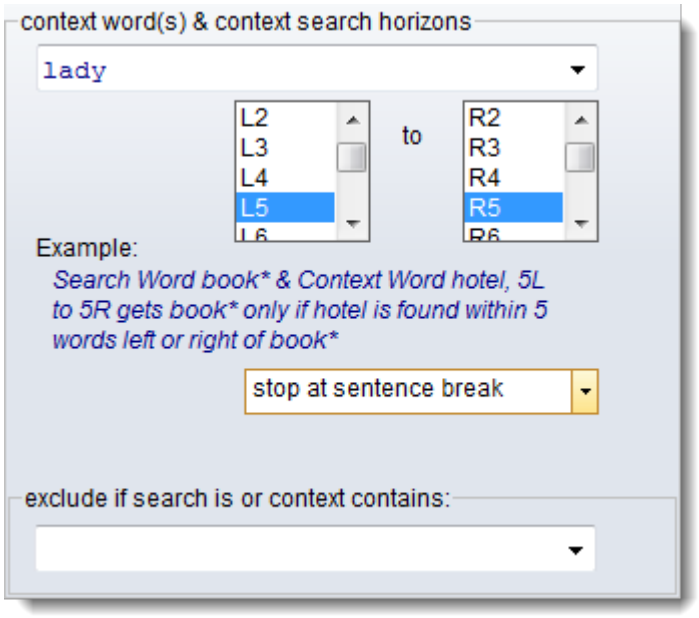


In this example the search-word is **beautiful** and the context word is **lady**, to be sought either left or right of **beautiful**.

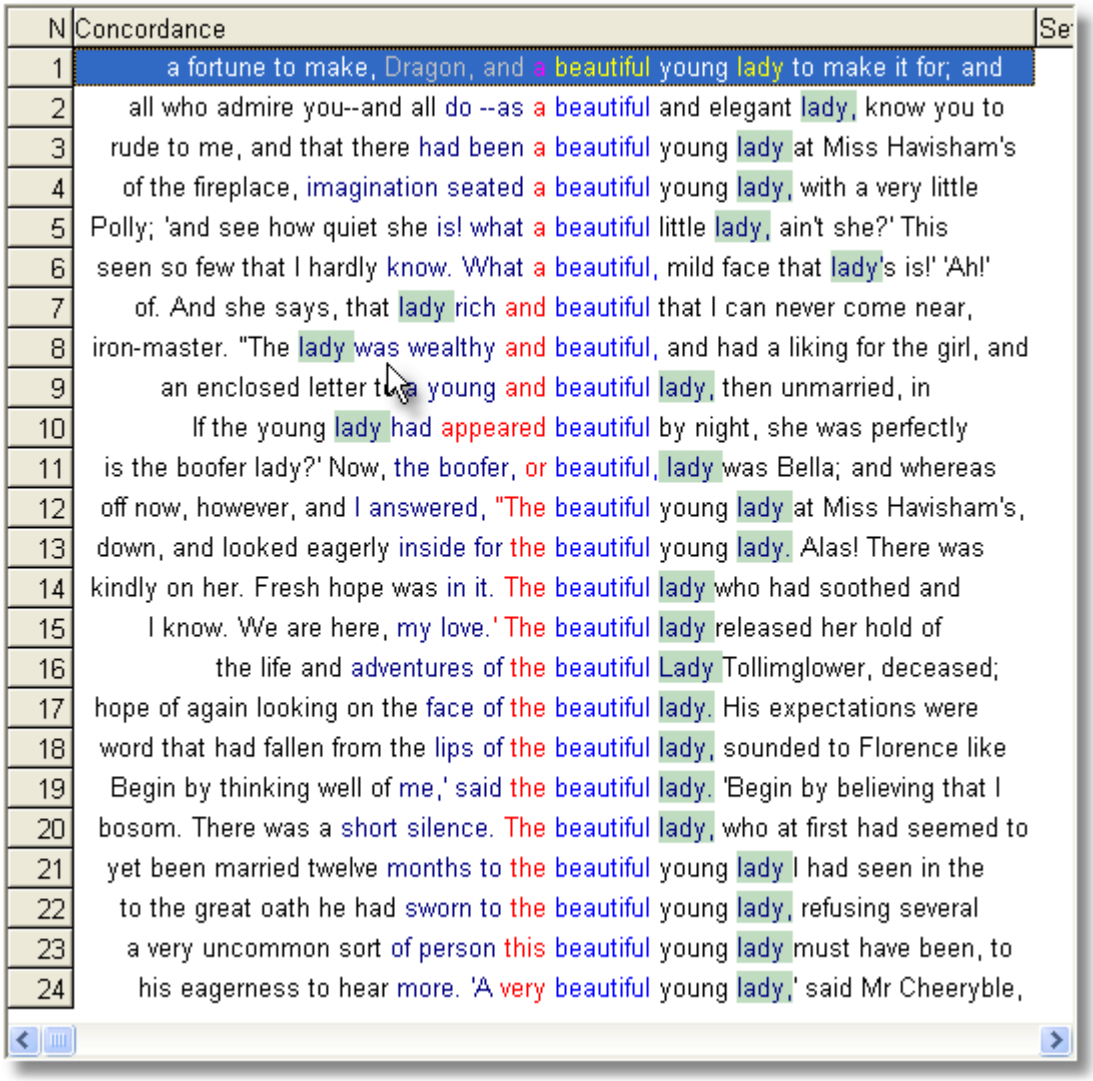
Syntax is like that of the search word or phrase 246,
* means disregard the end of the word and can be placed at either end of your context word.
== means case sensitive
/ separates alternatives. You can specify up to 15 alternatives within an 80-character overall limit.
If you want to use *, ? , == , ~ , :\ or / as a character in your search word, put it in double quotes, e.g. "*"

N	Concordance	Set	Ta
1	a fortune to make, Dragon, and a beautiful young lady to make it for; and		
2	all who admire you--and all do --as a beautiful and elegant lady, know you to		
3	rude to me, and that there had been a beautiful young lady at Miss Havisham's		
4	of the fireplace, imagination seated a beautiful young lady, with a very little		
5	Polly; 'and see how quiet she is! what a beautiful little lady, ain't she?' This		
6	seen so few that I hardly know. What a beautiful, mild face that lady's is!' 'Ah!'		
7	young lady who is good, accomplished, beautiful. You are good, accomplished,		
8	of. And she says, that lady rich and beautiful that I can never come near,		
9	iron-master. "The lady was wealthy and beautiful, and had a liking for the girl, and		
10	an enclosed letter to a young and beautiful lady, then unmarried, in		
11	If the young lady had appeared beautiful by night, she was perfectly		
12	is the boofer lady?' Now, the boofer, or beautiful, lady was Bella; and whereas		
13	full in Belinda's face, as he responded--'Beautiful, indeed!' The lady cast down		
14	not, my Lady Dedlock reigns supreme. Beautiful, elegant, accomplished, and		
15	off now, however, and I answered, "The beautiful young lady at Miss Havisham's,		
16	down, and looked eagerly inside for the beautiful young lady. Alas! There was		
17	kindly on her. Fresh hope was in it. The beautiful lady who had soothed and		
18	I know. We are here, my love.' The beautiful lady released her hold of		
19	the life and adventures of the beautiful Lady Tollimglower, deceased;		
20	hope of again looking on the face of the beautiful lady. His expectations were		
21	word that had fallen from the lips of the beautiful lady, sounded to Florence like		
22	Begin by thinking well of me,' said the beautiful lady. 'Begin by believing that I		
23	bosom. There was a short silence. The beautiful lady, who at first had seemed to		
24	yet been married twelve months to the beautiful young lady I had seen in the		
25	to the great oath he had sworn to the beautiful young lady, refusing several		
26	a very uncommon sort of person this beautiful young lady must have been, to		
27	his eagerness to hear more. 'A very beautiful young lady,' said Mr Cheeryble,		
28	is living now. She is a lady and very beautiful. And I love her!" With a last faint		

In line 14, the search-word and the context-word are in separate sentences. To avoid this, specify a suitable stop as shown here:



and with the same settings you will get results like these:



N	Concordance	Se
1	a fortune to make, Dragon, and a beautiful young lady to make it for; and	
2	all who admire you--and all do --as a beautiful and elegant lady, know you to	
3	rude to me, and that there had been a beautiful young lady at Miss Havisham's	
4	of the fireplace, imagination seated a beautiful young lady, with a very little	
5	Polly; 'and see how quiet she is! what a beautiful little lady, ain't she?' This	
6	seen so few that I hardly know. What a beautiful, mild face that lady's is!' 'Ah!'	
7	of. And she says, that lady rich and beautiful that I can never come near,	
8	iron-master. "The lady was wealthy and beautiful, and had a liking for the girl, and	
9	an enclosed letter t a young and beautiful lady, then unmarried, in	
10	If the young lady had appeared beautiful by night, she was perfectly	
11	is the boofer lady?' Now, the boofer, or beautiful, lady was Bella; and whereas	
12	off now, however, and I answered, "The beautiful young lady at Miss Havisham's,	
13	down, and looked eagerly inside for the beautiful young lady. Alas! There was	
14	kindly on her. Fresh hope was in it. The beautiful lady who had soothed and	
15	I know. We are here, my love.' The beautiful lady released her hold of	
16	the life and adventures of the beautiful Lady Tollimglower, deceased;	
17	hope of again looking on the face of the beautiful lady. His expectations were	
18	word that had fallen from the lips of the beautiful lady, sounded to Florence like	
19	Begin by thinking well of me,' said the beautiful lady. 'Begin by believing that I	
20	bosom. There was a short silence. The beautiful lady, who at first had seemed to	
21	yet been married twelve months to the beautiful young lady I had seen in the	
22	to the great oath he had sworn to the beautiful young lady, refusing several	
23	a very uncommon sort of person this beautiful young lady must have been, to	
24	his eagerness to hear more. 'A very beautiful young lady,' said Mr Cheeryble,	

If you have specified a context word, you can re-sort on it. Also, the context words will be in their own special colour 114.

Note: the search only takes place within the current concordance line with the number of characters defined as characters to save 242. That is, if for example you choose search horizons 25L and 25R, but only 1000 characters are saved in each line, there might not be 25 words on either side of the search-word to examine when seeking the context word or phrase if there was extensive mark-up as well.

7.10 dispersion plot

The point of it...

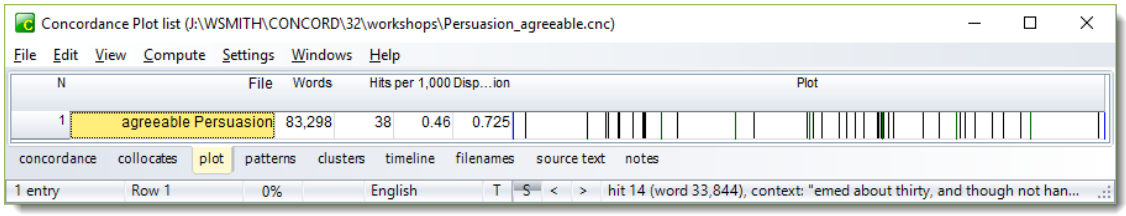
This shows where the search word occurs in the file which the current entry belongs to. That way you can see where mention is made most of your search word in each file. Another case where the aim is to promote the noticing of linguistic patterning 558.

What you see

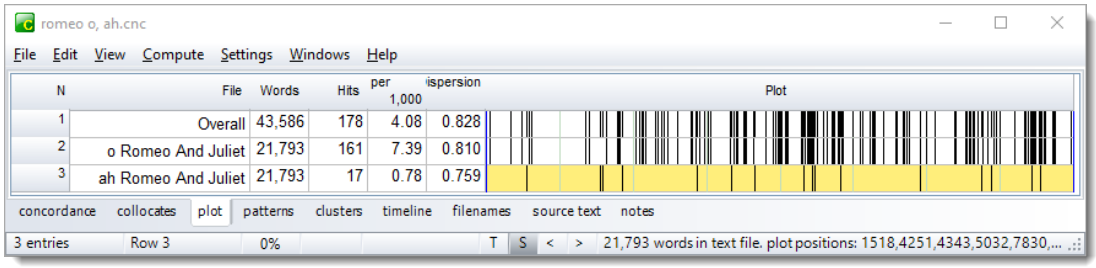
The plot shows:

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 554
Plot a plot showing where they cropped up, where the left edge of the plot represents the beginning of the text file ("Once upon a time" for example) and the right edge is at the end ("happily ever after". Though not in the case of Romeo and Juliet.).

If the settings 122 permit, Concord can compute the context of each hit and show it to you.



Here the mouse was hovering over the 14th vertical mark and the status bar shows the context. To move from one hit context to the next, use the > and < arrows in the status bar.



Here we see a plot of "O" and another of "AH" from the play Romeo and Juliet. They are on separate lines because there were 2 search-words. There are more "O" exclamations than "AH"s.

or double-click...

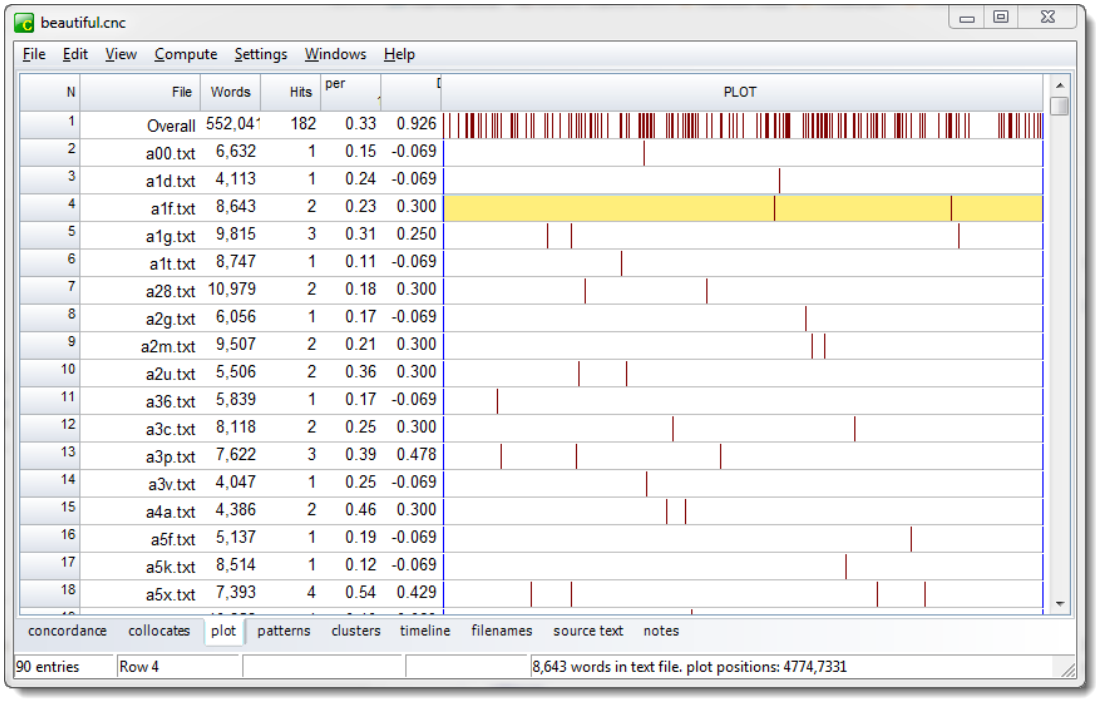
You can also double-click on a plot to see a list of all the word numbers 294.

Dispersion Details: ah Romeo And Juliet

N	plot positions
1	1518
2	4251
3	4343
4	5032
5	7830
6	9120
7	9949
8	11775
9	11784
10	12119
11	12197
12	13126
13	16113
14	18387
15	18745
16	20259
17	20583

Using *View* | *Ruler*, you can switch on a "ruler" splitting the display into segments.

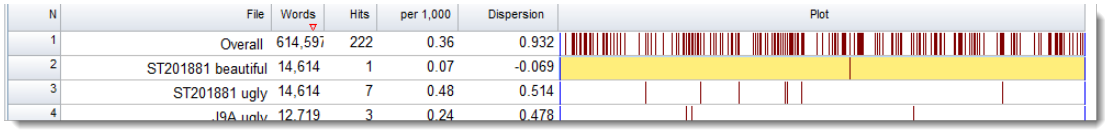
The plot below is of one search-word (**beautiful**) in lots of texts.



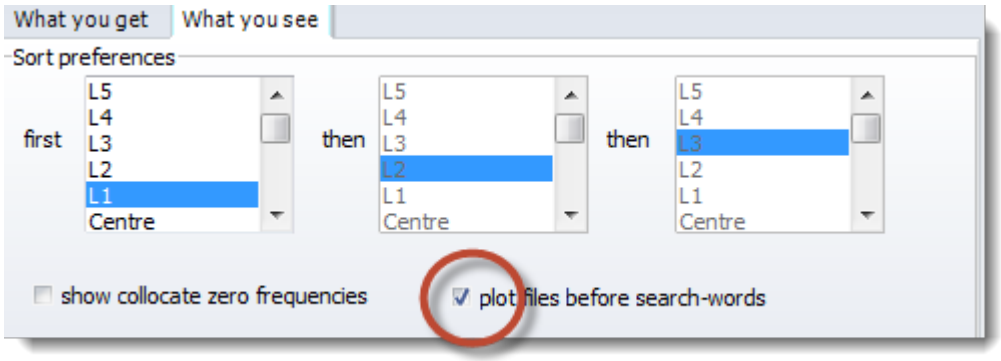
The status-bar gives details of the highlighted text.

Multiple Search-words or Texts

If there are 2 or more search-words or texts, you will see something like this:



where the *File* column supplies the file-name and the search-word in that order. If you want it with the search-word first, go to the Concord settings in the Controller, What you see, and click here:

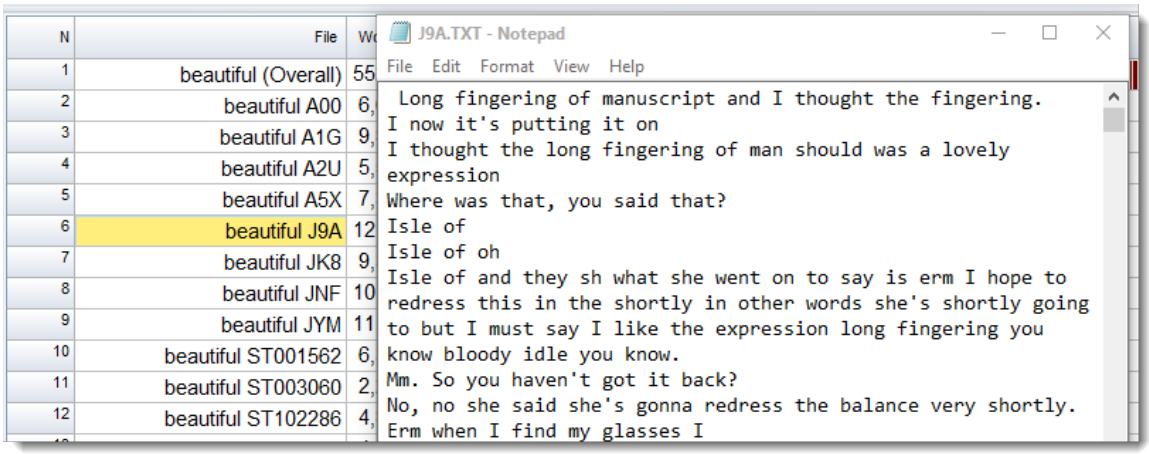


and re-sort the File list:

N	File	Words	Hits	per 1,000	ispersion
1	beautiful (Overall)	552,04	182	0.33	0.926
2	beautiful A00	6,632	1	0.15	-0.069
3	beautiful A1G	9,815	3	0.31	0.250
4	beautiful A2U	5,506	2	0.36	0.300
5	beautiful A5X	7,393	4	0.54	0.429
6	beautiful J9A	12,719	1	0.08	-0.069
7	beautiful JK8	9,395	4	0.43	0.192

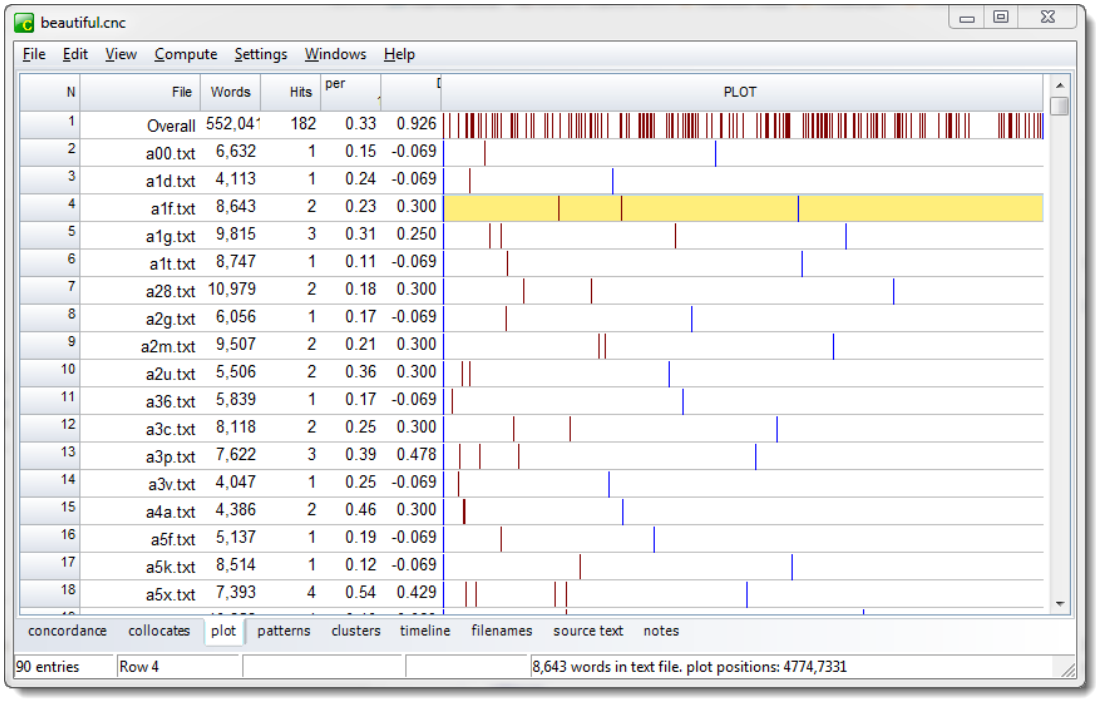
Double-click to see the source text

Just double-click in the File column:



Uniform view

There are two ways of viewing the plot, the default, where all plotting rectangles are the same length, or Uniform Plot (where the plot rectangles reflect the original file size -- the biggest file is longest). Change this in the View menu at the top. Here is the same one with Uniform plot. The blue edge at the right reflects the file size in each case.



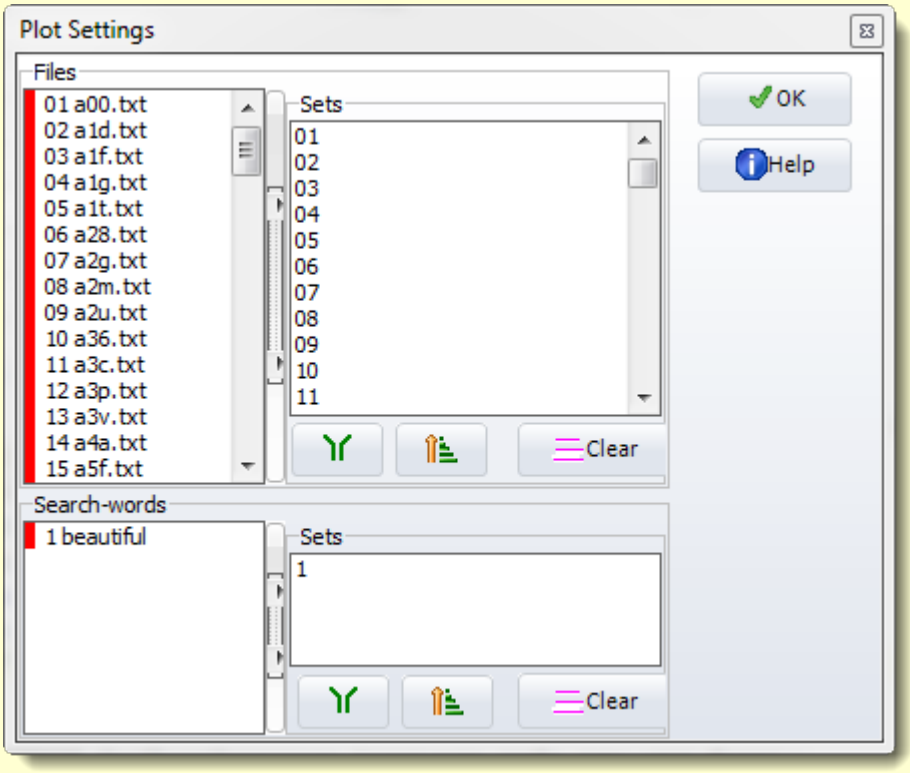
If you don't see as many marks as the number of hits, that'll be because the hits came too close together for the amount of screen space in proportion to your screen resolution. You can stretch the plot by dragging the top right edge of it. You can export the plot using Save As 91 and can get your spreadsheet to make graphs etc, as explained here 91.

Each plot window is dependent on the concordance from which it was derived. If you close the original concordance down, it will disappear. You can *Print* the plot. There's no *Save* option for the plot alone but you can of course save the concordance itself. You can *Copy* to the clipboard 529 (Ctrl+C) and then put it into a word processor as a graphic, using Paste Special.

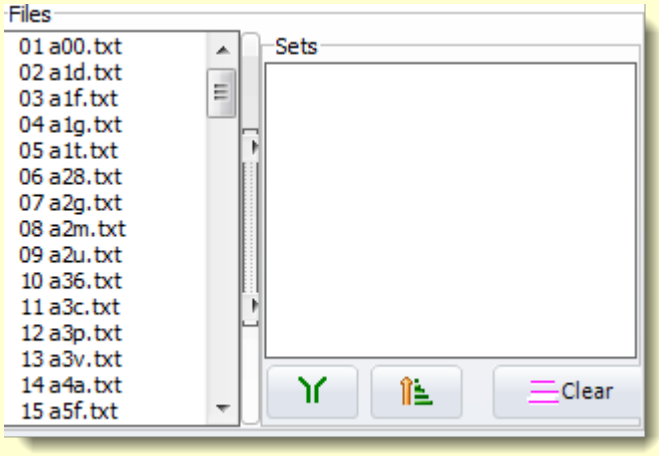
Advanced plots

When you first compute a concordance, the plot will assume you want a dispersion plot of each text file on a separate line and each different search-word on a separate line as seen above. If you have more than one text file or search-word, when you choose the *Compute* | *Plot* menu item afterwards, you will get a chance to merge your plots and omit some text files or search-words.

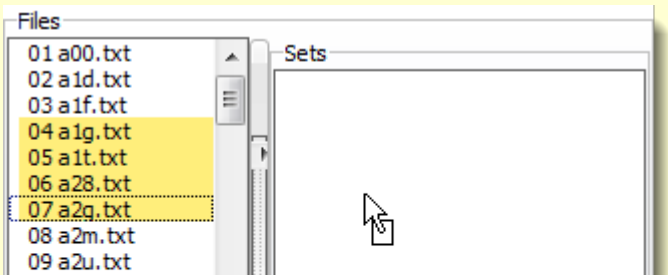
A first view of the plot settings may resemble this. All the files have by default been sorted into separate sets and so have all the search-words. The red colour indicates files or search-words which have been included in each list of sets at the right.



Now if you *Clear* them,



you can either select and drag or select and press the central button to get your preferred selections. (The button showing a green funnel will put all into one set, the other one will use one set for each, by the way.)



See also: plot and ruler colours 114, plot dispersion value 554, WordList dispersion statistic 354.

7.11 editing concordances

The point of it...

You may well find you have got some entries which weren't what you expected. Suppose you have done a search for **SHRIMP*/PRAWN*** -- you may find a mention of *Shrimpton* in the listing. It's easy to clean up the listing by simply pressing **Del** on each unwanted line. (Do a sort on the search word first so as to get all the *Shrimptons* next to each other.) The line will turn a light grey colour.

Pressing **Ins** will restore it, if you make a mistake. To delete or restore ALL the lines from the current line to the bottom, press the grey - key or the grey + key by the numeric keypad. When you have finished marking unwanted lines, you can choose (Ctrl+Z or) to zap 154 the deleted lines.

If you're a teacher you may want to blank 199 out the search words: to do so, press the spacebar. Pressing the spacebar again will restore it, so don't worry!

7.11.1 remove duplicates

The problem

Sometimes one finds that text files contain duplicate sections, either because the corpus has become corrupted through being copied numerous times onto different file-stores or because they were not edited effectively, e.g. a newspaper has several different editions in the same file. The result can sometimes be that you get a number of repeated concordance lines.

Solution

If you choose *Edit |Remove Duplicates*, **Concord** goes through your concordance lines and if it finds any two where the concordance lines match each other (regardless of the filename, date etc.) it will mark one of these for deletion. To establish a match it examines 500 characters centred on the search-word in each line. Every single character including punctuation would need to be identical for a match.

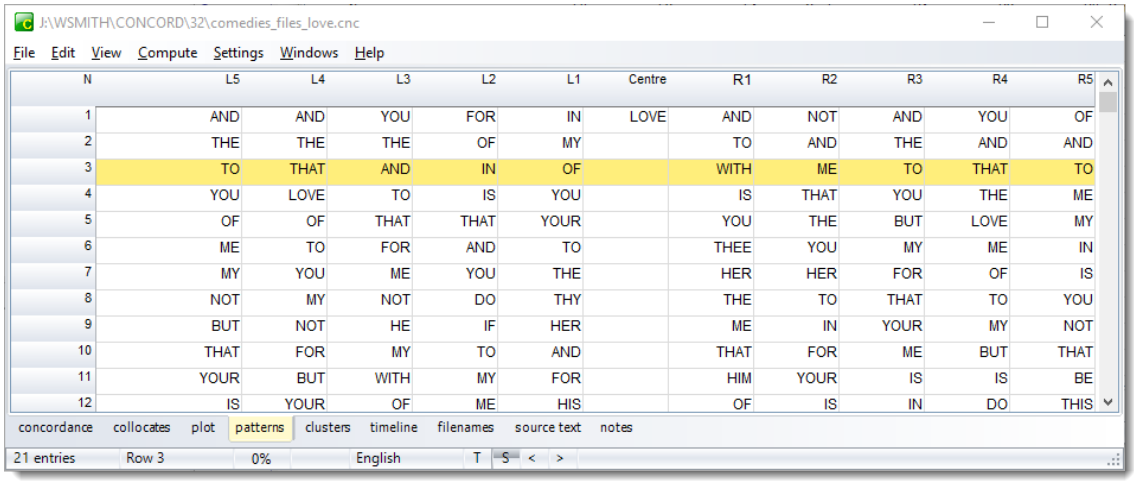
Check before you zap...

At the end it will sort all the lines so you can see which ones match each other before you decide finally to zap 154 the ones you really don't want.

See also: the Corpus Checker utility's find duplicate contents 422 function which finds near duplicates.

7.12 patterns

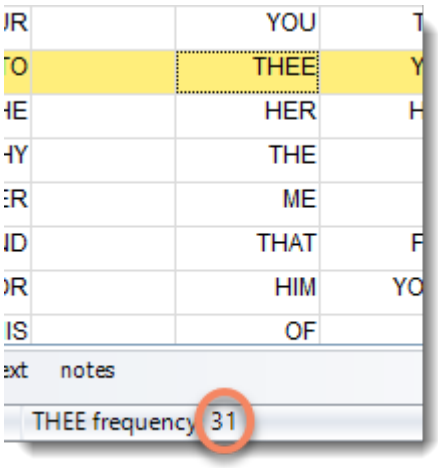
When you have a collocation window open, one of the tab windows shows "Patterns". This will show the collocates (words adjacent to the search word), organised in terms of frequency within each column. That is, the top word in each column is the word most frequently found in that position. The second word is the second most frequent.



N	L5	L4	L3	L2	L1	Centre	R1	R2	R3	R4	R5
1	AND	AND	YOU	FOR	IN	LOVE	AND	NOT	AND	YOU	OF
2	THE	THE	THE	OF	MY		TO	AND	THE	AND	AND
3	TO	THAT	AND	IN	OF		WITH	ME	TO	THAT	TO
4	YOU	LOVE	TO	IS	YOU		IS	THAT	YOU	THE	ME
5	OF	OF	THAT	THAT	YOUR		YOU	THE	BUT	LOVE	MY
6	ME	TO	FOR	AND	TO		THEE	YOU	MY	ME	IN
7	MY	YOU	ME	YOU	THE		HER	HER	FOR	OF	IS
8	NOT	MY	NOT	DO	THY		THE	TO	THAT	TO	YOU
9	BUT	NOT	HE	IF	HER		ME	IN	YOUR	MY	NOT
10	THAT	FOR	MY	TO	AND		THAT	FOR	ME	BUT	THAT
11	YOUR	BUT	WITH	MY	FOR		HIM	YOUR	IS	IS	BE
12	IS	YOUR	OF	ME	HIS		OF	IS	IN	DO	THIS

In R1 position (one word to the right of the search-word **love**) there seem to be both intimate (**thee**) and formal (**you**) pronouns associated with **love** in Shakespeare's comedies. And looking at L1 position it seems that speakers talk more of their love for another than of another's love for them.

The minimum frequency and length for one of the words to be shown at all, is the minimum frequency/length for collocates 217. You can also select a word to see its frequency in the status bar:



The point of it…

The effect is to make the most frequent items in the neighbourhood of the search word "float up" to the top. Like collocation, this helps you to see lexical patterns in the concordance.

You can also highlight any given pattern collocate 215 in your concordance display.

7.13 re-sorting

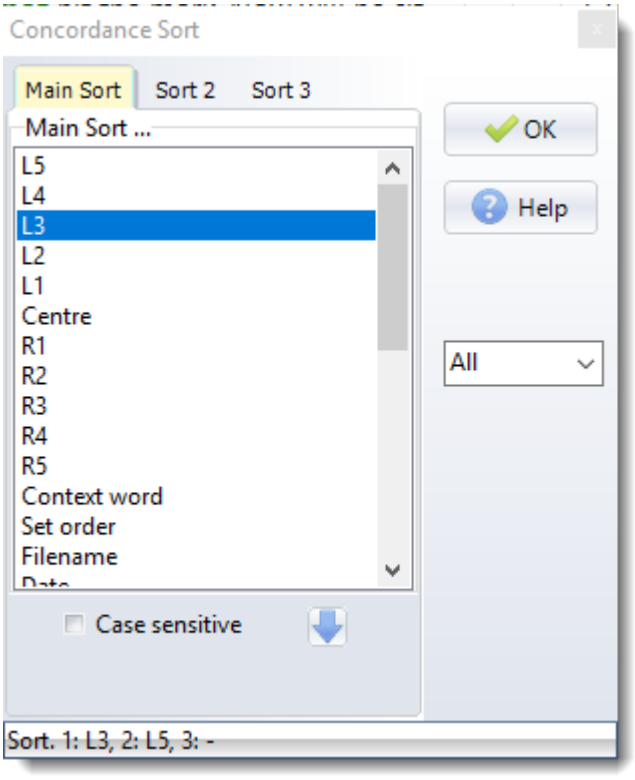
How to do it...

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

The point of it…

The point of re-sorting is to find characteristic lexical patterns. It can be hard to see overall trends in your concordance lines, especially if there are lots of them. By sorting them you can separate out multiple search words and examine the immediate context to left and right. For example you may find that most of the entries have "in the" or "in a" or "in my" just before the search word -- sorting by the second word to the left of the search word will make this much clearer.
Sorting is by a given number of words to the left or right (L1 [=1 word to the left of the search word], L2, L3, L4, L5, R1 [=1 to the right], R2, R3, R4, R5), on the search word itself, the context word (if one was specified), the nearest tag 223, the distance to the nearest tag, a set category 43 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 (L1) will sort the entries according to the alphabetical order of the word immediately to the left of the search word. A second sort (Sort 2) on R2 would re-order the listing by tie-breaking, that is: only where the L1

words (immediately to the left of the search word) matched exactly, and would place these in alphabetical order of the words 2 to the right of the search word. For very large concordances you may find the third sort (Sort 3) useful: this is an extra tie-breaker in cases where the second sort matches.
For many purposes tie-breaking is unnecessary, and will be ignored if the "activated" box is not checked.

default sort

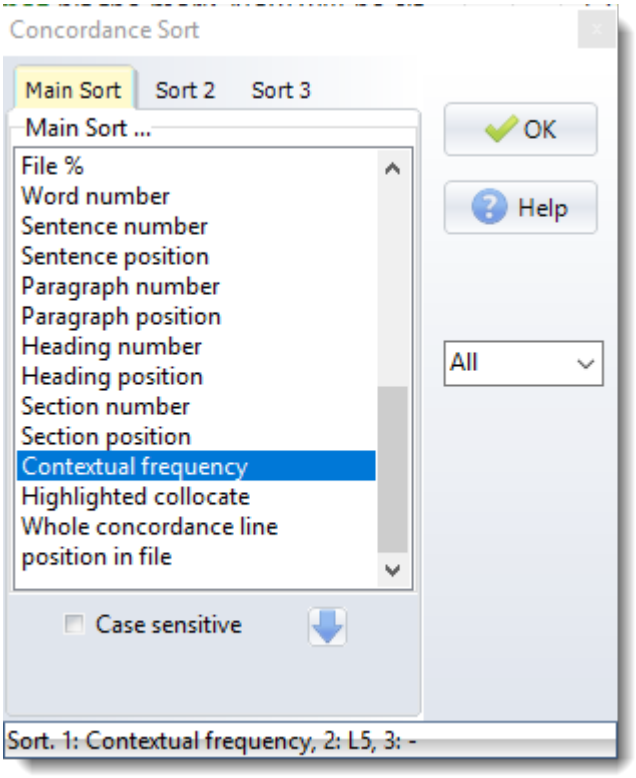
This is set in the main controller settings 257.

sorting by set (user-defined categories 43)

You can also sort by set, if you have chosen to classify the concordance lines according to your own scheme, using letters from **A** to **Z** or **a** to **z** or longer strings. The sort will put the classified lines first, in category order, followed by any unclassified lines. See Nearest Tag 223 for details of sorting by tags. See colour categories 35 for a more sophisticated way of using the Set column.

other sorts

As the screenshot below shows, you can also sort by a number of other criteria, most of these accessible simply by clicking on their column header.



The "contextual frequency" sort means sorting on the average ranking frequency of all the words in each concordance line which don't begin with a capital letter. For this you will be asked to specify your reference corpus 555 wordlist. The result will be to sort those lines which contain "easy" (highly frequent) words at the top of the list.

All

By default you sort all the lines; you may however type in for example 5-49 to sort those lines only.

Ascending or Descending?

Use the **or** button to set the sort order is from **A** to **Z**, or **Z** to **A**.

See also: WordList sort 372, KeyWords sort 296, Choosing Language 77

7.13.1 re-sorting: dispersion plot

This automatically re-sorts the dispersion plot, rotating through these options:
alphabetically (by file-name)
in *frequency* order (in terms of hits per 1,000 words of running text)
by first occurrence in the source text(s): *text order*
by *range*: the gap between first and last occurrence in the source text.

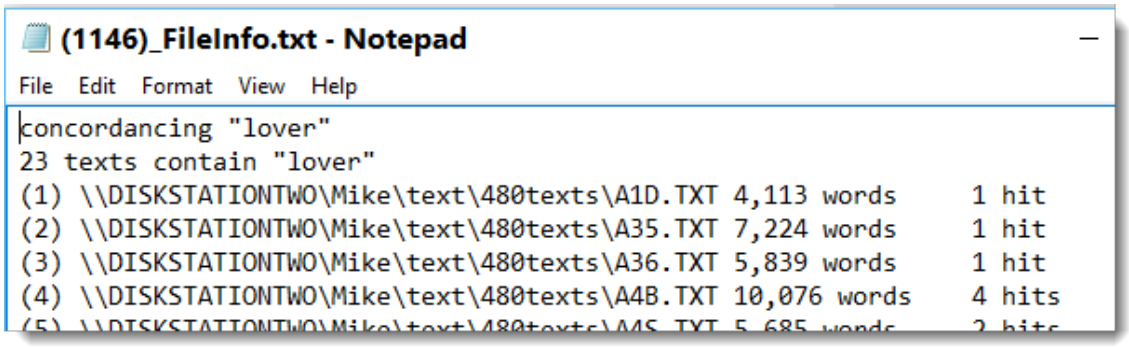
see also: Dispersion Plot 230

7.14 file-names data

The point of it

When you have computed a concordance you may wish to know which text files contained and which did not contain any of your search-words. (The picture might get altered if you delete and zap some concordance lines.)

This function gives you a plain text list.

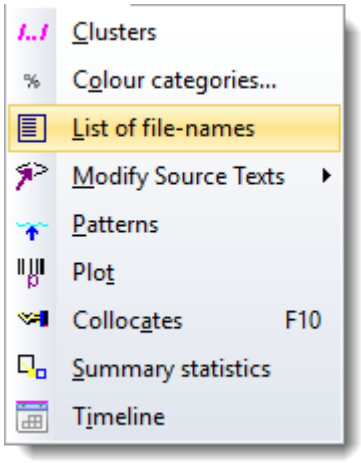


These data can be exported to Excel if so desired.

How to do it

To generate the list automatically see Advanced settings 122: save text file data.

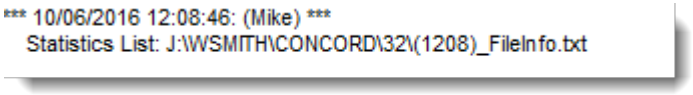
Manually: start the process using the *Compute* | *List of file-names* menu option.



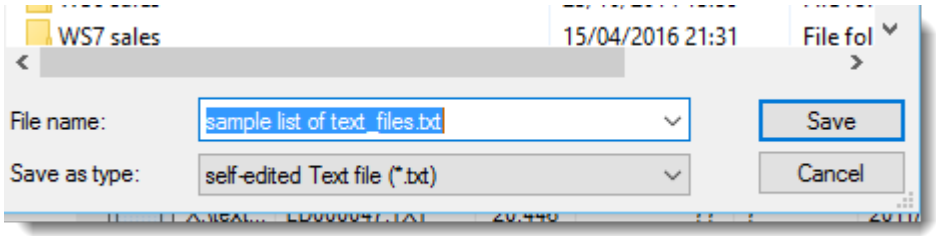
You get a list showing first all the texts which did contain any of your search-words, then below a list of those which did not.

```
(23) \\DISKSTATIONTWO\Mike\text\480texts\ST202376.NEW   2,550 word
457 texts do not contain "lover"
(1) \\DISKSTATIONTWO\Mike\text\480texts\A00.TXT 6,632 words
(2) \\DISKSTATIONTWO\Mike\text\480texts\A01.TXT 7,826 words
(3) \\DISKSTATIONTWO\Mike\text\480texts\A0A.TXT 5,732 words
(4) \\DISKSTATIONTWO\Mike\text\480texts\A0W.TXT 10,338 words
(5) \\DISKSTATIONTWO\Mike\text\480texts\A13.TXT 4,376 words
```

And a note gets added to the Notes tab showing the name of the text statistics file created.



If you use data created in earlier versions, you will be asked to supply a list of all the text files used for that concordance data. If you haven't got one, create a suitable list using the file-choose favourites 59 function: ensure you choose to save the selected texts as *self-edited Text file* **`.txt`** format.



See also: re-compute file-name lists 242, statistics 350, auto-saving file data 122

7.14.1 recompute file lists

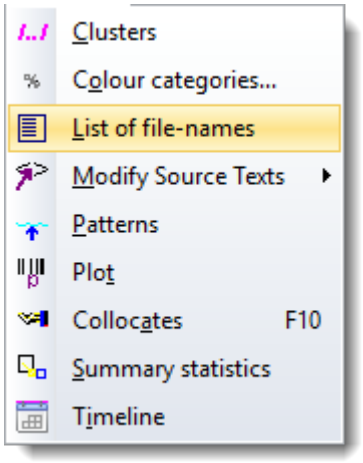
Choose a matching .cnc file ...

You might find you are not getting successful connections to the source text file 200 after a lot of editing if the file-names have got misaligned.

If Concord detects a mismatch in the file data, you may get a warning reporting a file mismatch, and an automatic request to supply a **.cnc** file with trustworthy data. Simply compute another **.cnc** on the same text files using the same search-term(s). Do not trouble to edit this but save it. The saved **.cnc** file will now serve as a trustworthy source to fix links in the previous edited **.cnc** file and the connections to the source texts should be restored.

How to do it manually

The menu option to make this happen manually is *Compute* | *List of file-names*.



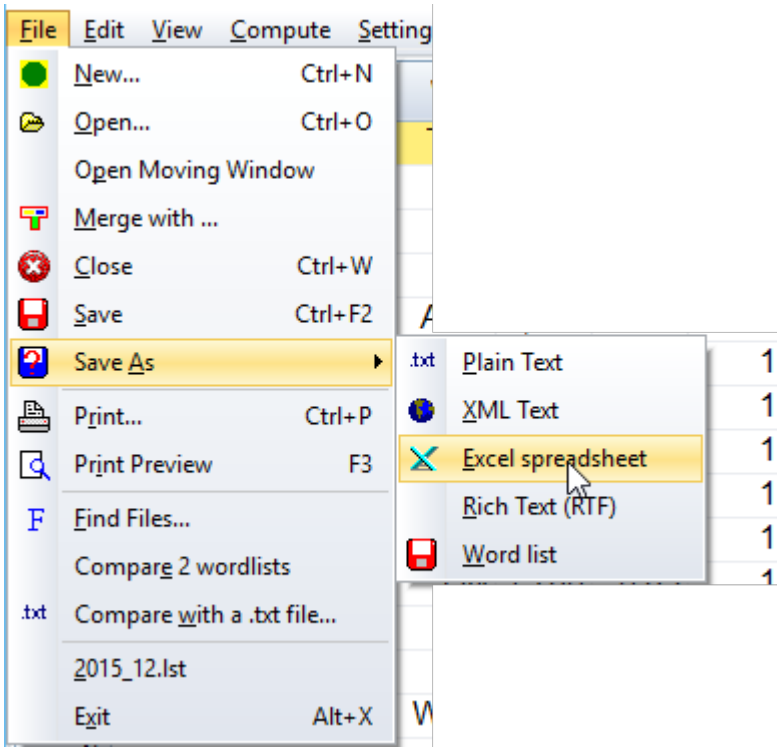
See also: file names data 240, viewing the source text file 200.

7.15 saving and printing

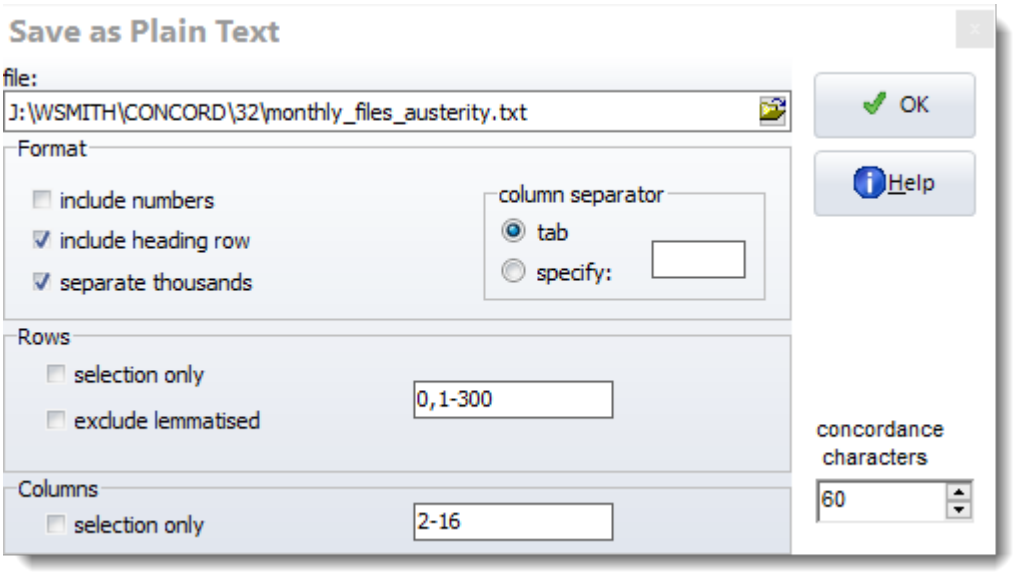
You can save the concordance (and its collocates & other dependent results if these were stored when the concordance was generated) either as a Text File (e.g. for importing into a word processor) or as a file of results which you can subsequently *Open* (in the main menu at the top) to view again at 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 save as in various formats:



Example

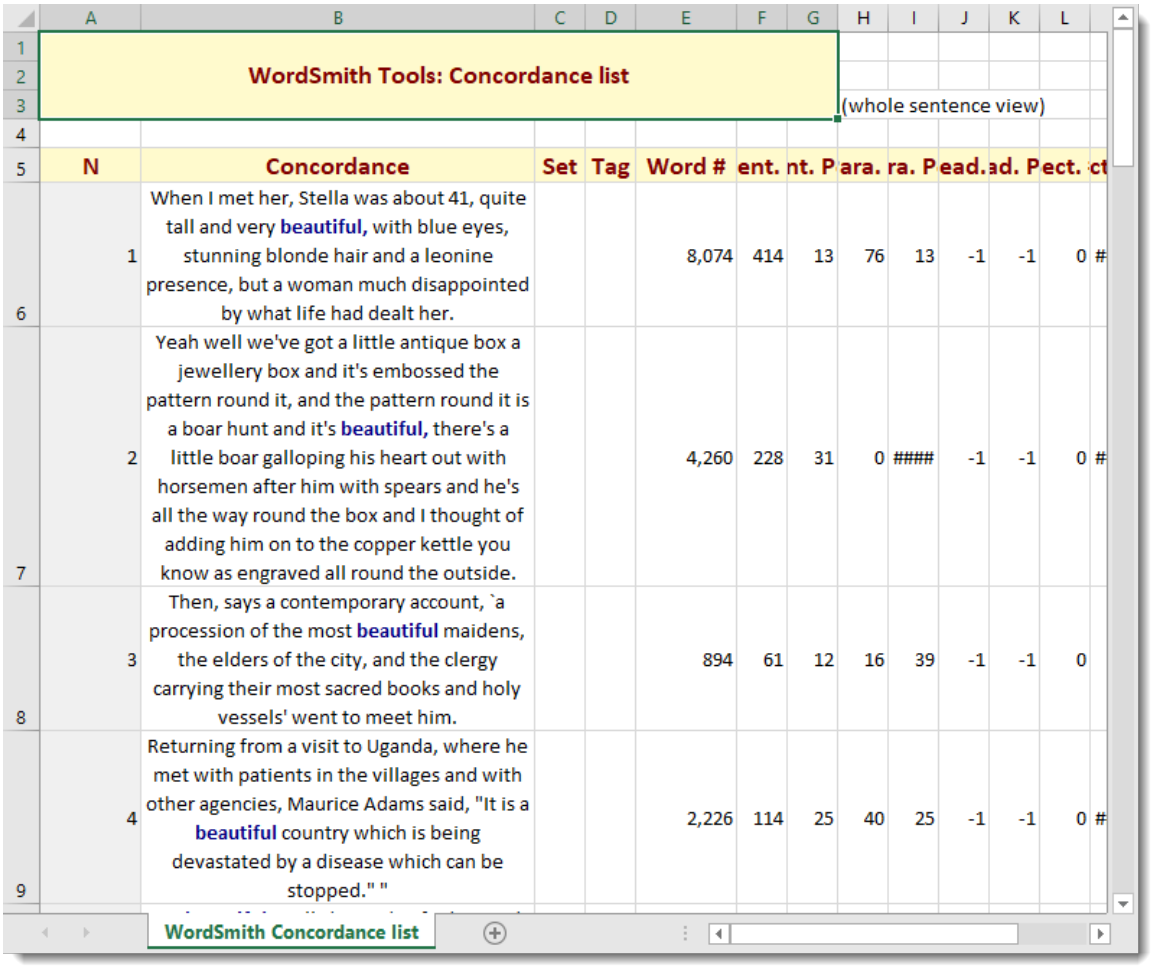


These options saved the header row plus 300 lines with all the concordance columns except the number column at the left, and in each concordance line only 60 characters were saved.

Concordance characters

The idea is to save only a fixed number of characters per concordance line. That makes it easy to display the lines nicely lined up e.g. in a thesis or in teaching materials.

If you set *concordance characters* to 0 you will get the whole concordance length exported, or if you you have set the view to sentence only [197] you will just get the sentence exported:



WordSmith Tools: Concordance list (whole sentence view)

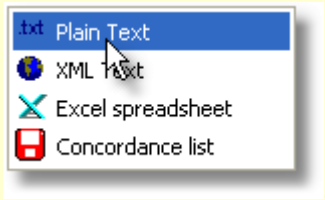
N	Concordance	Set	Tag	Word #	ent.	nt.	Para.	ra.	Read.	ad.	Pect.	ct
1	When I met her, Stella was about 41, quite tall and very **beautiful,** with blue eyes, stunning blonde hair and a leonine presence, but a woman much disappointed by what life had dealt her.			8,074	414	13	76	13	-1	-1	0	#
2	Yeah well we've got a little antique box a jewellery box and it's embossed the pattern round it, and the pattern round it is a boar hunt and it's **beautiful,** there's a little boar galloping his heart out with horsemen after him with spears and he's all the way round the box and I thought of adding him on to the copper kettle you know as engraved all round the outside.			4,260	228	31	0	####	-1	-1	0	#
3	Then, says a contemporary account, `a procession of the most **beautiful** maidens, the elders of the city, and the clergy carrying their most sacred books and holy vessels' went to meet him.			894	61	12	16	39	-1	-1	0	
4	Returning from a visit to Uganda, where he met with patients in the villages and with other agencies, Maurice Adams said, "It is a **beautiful** country which is being devastated by a disease which can be stopped." "			2,226	114	25	40	25	-1	-1	0	#

WordSmith Concordance list

Print

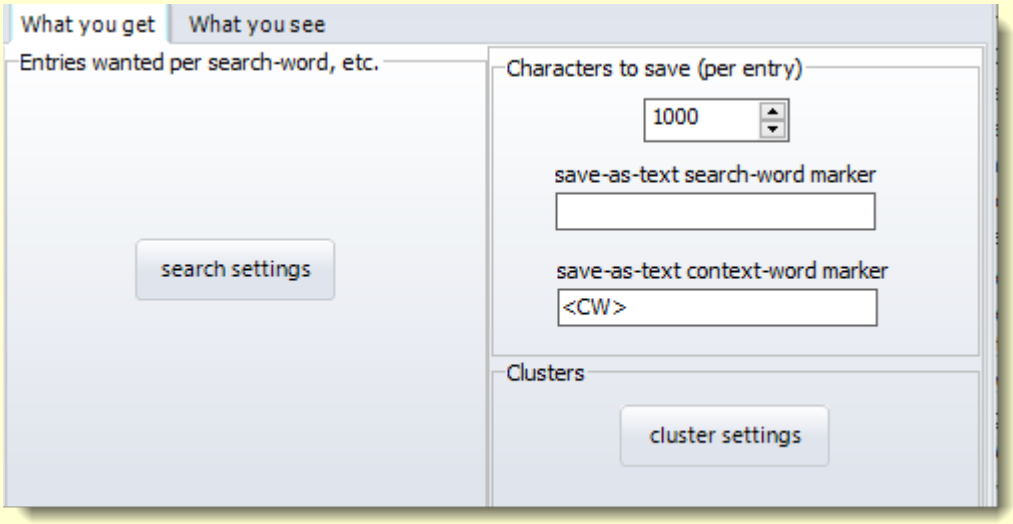
You can print [113] using any printer attached to your system. You will get a chance to specify the number of pages to print. The font will approximate the one you can see on your screen. If you use a colour printer or one with various shades of grey, the screen colours will be copied to your printer. 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 [88].

Insert search-word or context-word markup

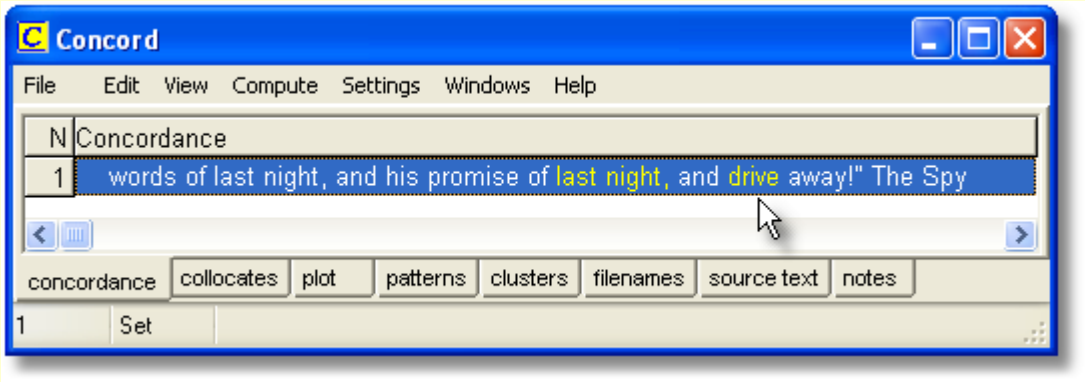


If you choose to save as text using , and if you have (optionally) marked out the search-word and/or context word in the Controller 253

like this



whatever you have put will get inserted in the **.txt** file. In the above example, doing a search through 23 Dickens texts for **last night** with **drive** as the context word, a concordance looking like this



produced this in the txt file:

rry, tell him yourself to give him no restorative but air, and to remember my words of last night, and his promise of last night, and <CW>drive away!" The Spy withdrew, and Carton seated himself at the table, resting his forehead on his h

See also : using the clipboard 529 to get a concordance into Word or Excel another application.

7.16 search-word or phrase

7.16.1 search word syntax

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

Basic Examples

search word	finds
book	*Book* or *book* or *BoOk*
book*	*book, books, booking, booked*
***book**	*textbook* (but not *textbooks*)
b*	*banana, baby, brown etc.*
***ed**	*walked, wanted, picked etc.*
bo* in	*book in, books in, booking in* (but not *book into*)
book * hotel	*book a hotel, book the hotel, book my hotel*
bo* in*	*book in, books in, booking in, book into*
book?	*book, books, book; book.*
book^	*book, books*
b^^k	*book, back, bank,* etc.
==book==	*book* (but not *BOOK* or *Book*)
book/paperback	*book* or *paperback*

symbol	meaning	examples
*	disregard the end of the word,	**tele***
	disregard a whole word	***ness**
		happi
		book * hotel
?	any single character (including punctuation) will match here	**Engl???**
		?50.00
#	any sequence of numbers, 0 to 9	**$#**
		£#.00
^	any single letter of the alphabet will match here	**Fr^nc^**
==	case sensitive	**==French==**
		==Fr*==
/\	end of line	**/\Sentence**
		line end./
:\	means use a file for lots of search-words (see file-based	**c:\text\frd.txt**

search_words 247)

/	separates alternative search-words. You can specify alternatives within an 80-character overall limit	**may/can/will**
<>	beginning & end of tags	**<w NN1>**

Advanced Search-word Syntax

If you want to use ***, ? , == , #, ^ , :\, >, <** or **/** as a character in your search word, put it in double quotes. Examples:

"*"
Why"?"
and"/"or
":\"
"<"

Don't forget that question-marks come at the end of words (in English anyway) so you might need ***"?"**

If you need to type in a symbol using a code, they can be symbolised like this: **{CHR(xxx)}** where **xxx** is the decimal number of the code. Examples: **{CHR(13)}** is a carriage-return, **{CHR(10)}** is a line-feed, **{CHR(9)}** is a tab. To represent **<Enter>** which comes at the end of paragraphs and sometimes at the end of each line, you'd type **{CHR(13)}{CHR(10)}** which is carriage-return followed immediately by line-feed.

{CHR(34)} refers to double inverted commas.

{CHR(46)} is a full stop. There is a list of codes at http://www.asciitable.com/

You can also use hex format for numbers, e.g. **#x9** for tab, **#x22** for double inverted commas.

Tags

You can also specify tags in your search-word if your text is tagged. For XML formats see XML text handling 189

See also: Tag Concordancing 221, Context Word 226, Modify source texts 182, Ignore punctuation 255, Wildcards 148

7.16.2 file-based search-words

The point of it…

To save time typing in complex searches.

You may want to do a standard search repeatedly on different sub-corpora.

Or as Concord allows an almost unlimited number of entries, you may wish to do a concordance involving many search-words or phrases 246.

The space for typing in multiple search-words is limited to 80 characters (including / etc.). If your preferred search-words will exceed this limit or you wish to use a standardised search, you can prepare a file containing all the search-words.

How to do it...

A sample (**`Documents\wsmith7\concordance_search_words.txt`**) is included with the distribution files.

Use a Windows editor (e.g. Notepad) to prepare your own. Each one must be on a separate line of your file. No comment lines can be included, though blank lines may be inserted for readability.

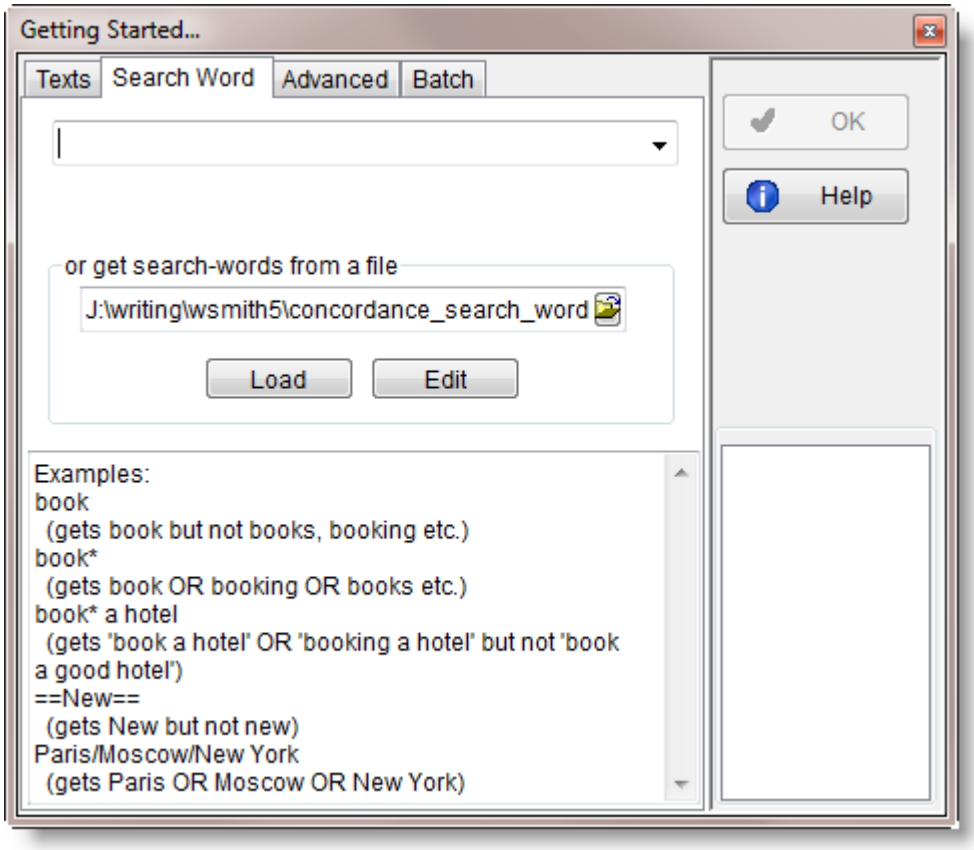
If you want to require a context for a given word, put **`context:=`** as in this example:

```
book context:=hotel
```

(which seeks **book** and only shows results if **`hotel`** comes in the context horizons).

Then, instead of typing in each word or phrase in the Search Word dialogue box, just browse for the file.

Then press *Load* to read the entries (or *Clear* if you change your mind).



concordancing from WordList, KeyWords etc.

Where Concord has been called up[545] from WordList, and the highlighted word in the word list is the head entry with lemmas[322], 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.

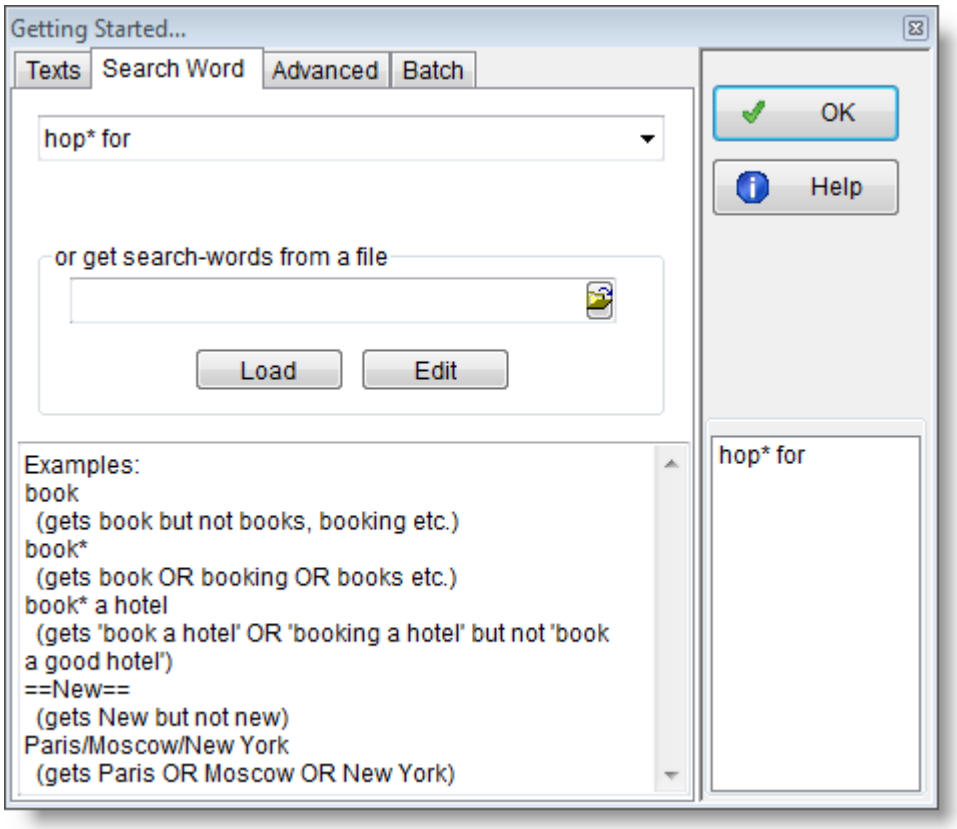
Automated file-based concordances

If you want Concord to process a whole lot of different search-words, saving each result as it goes along so you can get a lot of work done with WordSmith unattended, choose *SW Batch* under Concordance Settings[249].

7.16.3 search-word and other settings

Search Word or Phrase and/or Tag

Type the word or phrase[246] Concord will search for when making the concordance, or (below) the name of a file of search words[247]. You may also choose from a history list[543] of your previous search words. For details of syntax, see Search Word Syntax[246] or the set of examples shown in this screenshot:



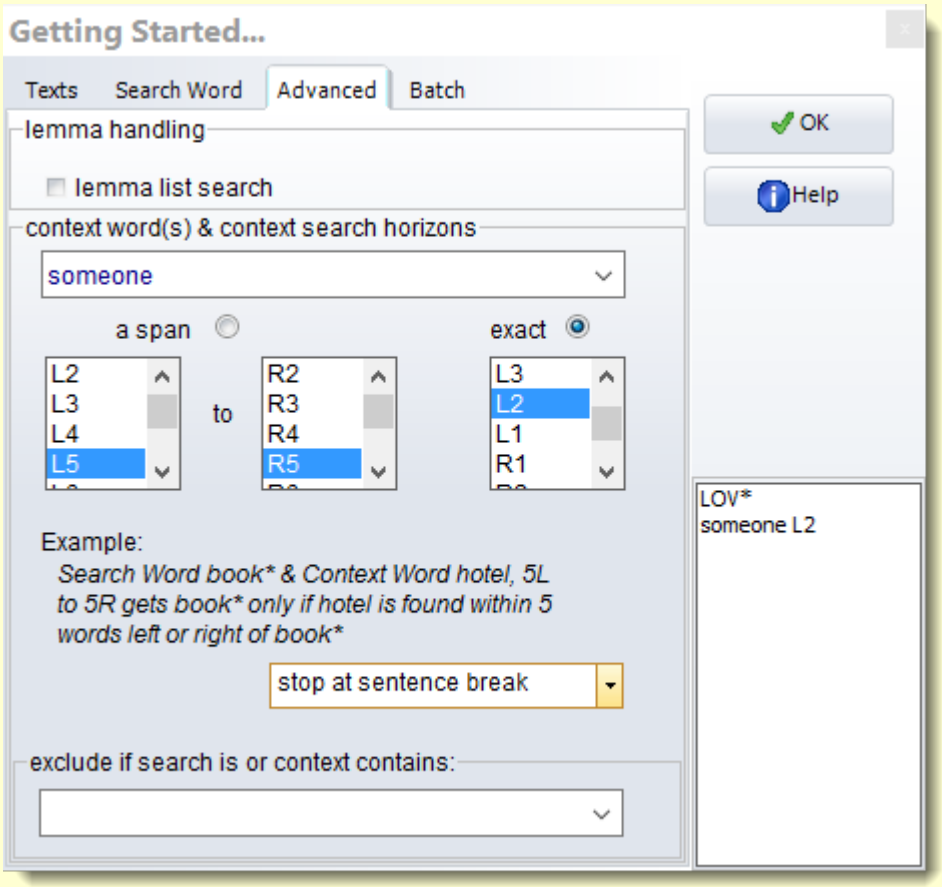
If you want to do many concordances in a file-based search[247], first prepare a small text file containing the search words, e.g. containing

```
this
```

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

Press the file button to locate your text file, the press the *Load* button. This will then change its name to something like *Clear 4*, where 4 means as in the example above that there are 4 different search-words to be concordanced. See "Batch" below for details on saving each one under a separate filename, otherwise all the searches will be combined into the same concordance.

Advanced & Batch searches



lemma list search

This option requires you to have chosen and loaded a lemma file[327]. If the lemma file you've loaded specifies for example **speak -> speaks, spoke, spoken** then if your search-word is **speak**, the concordance will contain examples of all four forms.

Context word(s) and search horizons

You may wish to find a word or phrase depending on the context. In that case you can specify context word(s) which you want, or which you do not want (and if found will mean that entry is not used).

For example, if the search word is **book*** and the context word is **hotel**, you'll get **book, books, booked, booking, bookable**, but only if **hotel** is found within your Context Search Horizons[226].

The search span is either a range such as L5 up to R5, or else an exact position. In the shot here the exact L2 position has been chosen, requiring the search-word **lov*** to have **someone** exactly 2 positions to its left.

You may type tag mark-up in here too, e.g. search for **book** with a context word **<ADJ>*** in position up to L3 will find book with a preceding adjective if your text has that sort of mark-up and if you've defined a tag file including **<ADJ>.**

In the screenshot above you see that "stop at sentence break" has been selected, meaning that a collocation search will only go left or right of the search-word up to a sentence-end. This is further explained here[217].

exclude if search is or context contains:

If the search word is **book*** and the *exclude if* box has **hotel**, you'll get **book, books, booked, booking, bookable**, as long as **hotel** is *not* found within your context search horizons. Or if the search word is ***ish** and the exclusion specifies **fish**, you'll get **yellowish, greenish**, etc. but not **fish**. You can insert more than one item, e.g. **fish/squish.**

Batch

Other settings affecting a concordance are available too:
see WordSmith Controller Concordance Settings 253; Typing characters 528, Accented characters 527; Choosing Language 77, Context Word(s) & Context Search Horizons 226

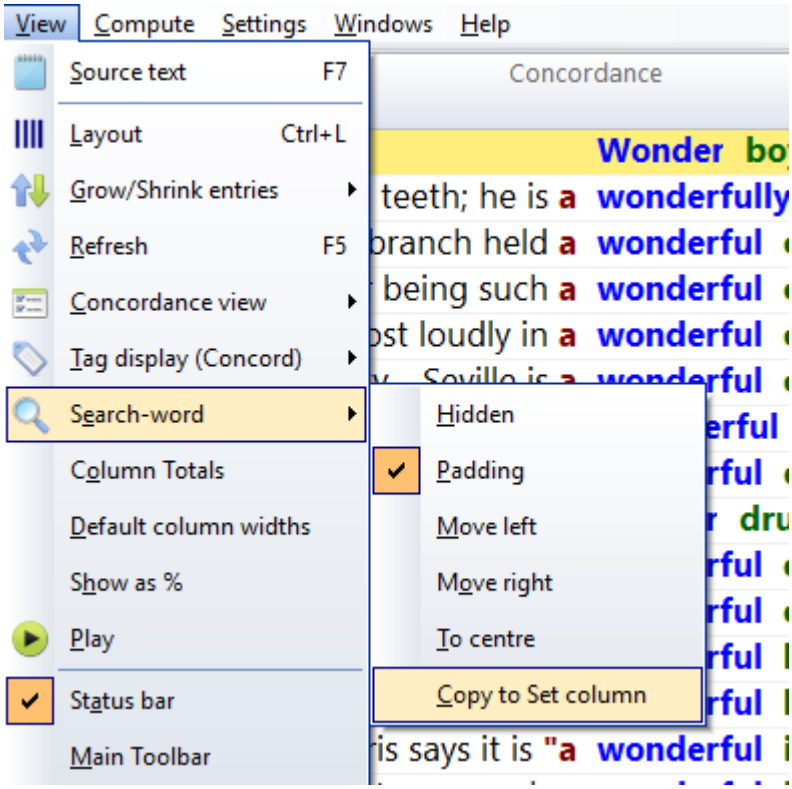
7.16.4 copy search-word to Set column

The point of it

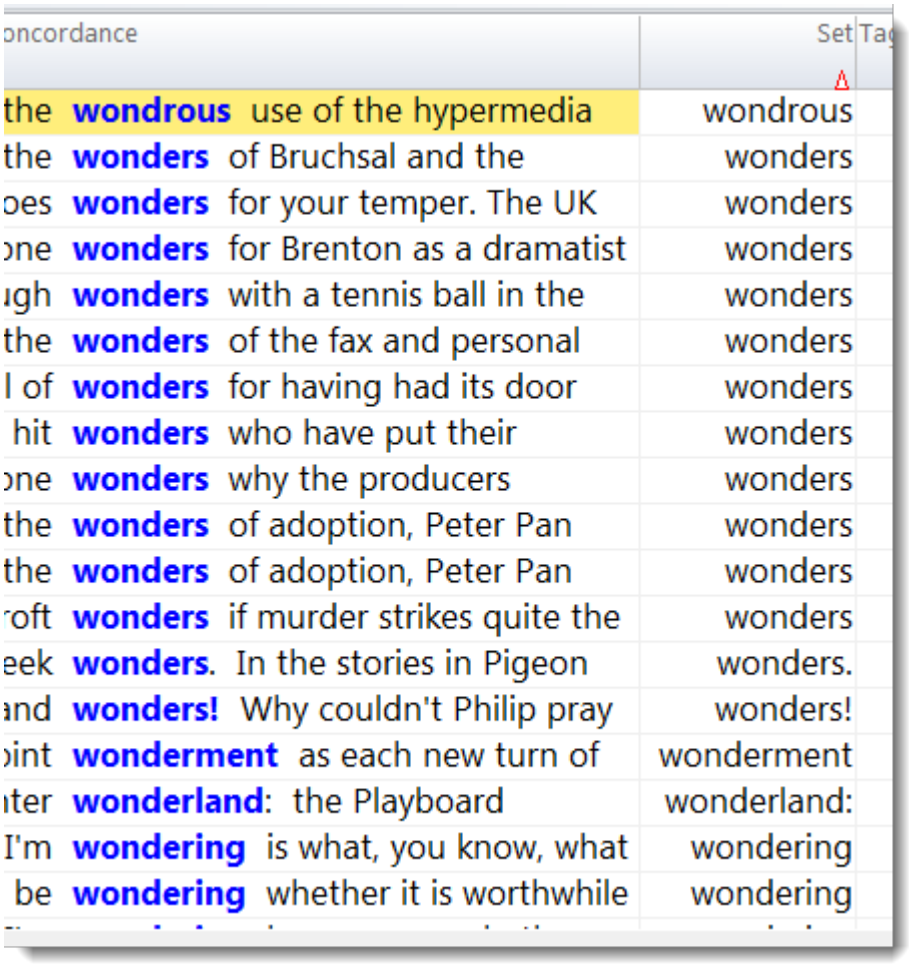
The idea is to be able to see a list of the search-word forms and for example export the list to a spreadsheet. Useful if you are using mixed search-words in the same concordance.

How to do it

With a concordance on screen, choose View | Search-word | Copy to Set column.



and you get something like this, where each search-form is shown in the Set column.

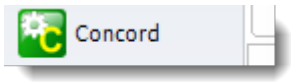


oncordance	Set Tag
the **wondrous** use of the hypermedia	wondrous
the **wonders** of Bruchsal and the	wonders
oes **wonders** for your temper. The UK	wonders
ne **wonders** for Brenton as a dramatist	wonders
gh **wonders** with a tennis ball in the	wonders
the **wonders** of the fax and personal	wonders
l of **wonders** for having had its door	wonders
hit **wonders** who have put their	wonders
ne **wonders** why the producers	wonders
the **wonders** of adoption, Peter Pan	wonders
the **wonders** of adoption, Peter Pan	wonders
oft **wonders** if murder strikes quite the	wonders
eek **wonders**. In the stories in Pigeon	wonders.
nd **wonders!** Why couldn't Philip pray	wonders!
int **wonderment** as each new turn of	wonderment
ter **wonderland**: the Playboard	wonderland:
I'm **wondering** is what, you know, what	wondering
be **wondering** whether it is worthwhile	wondering

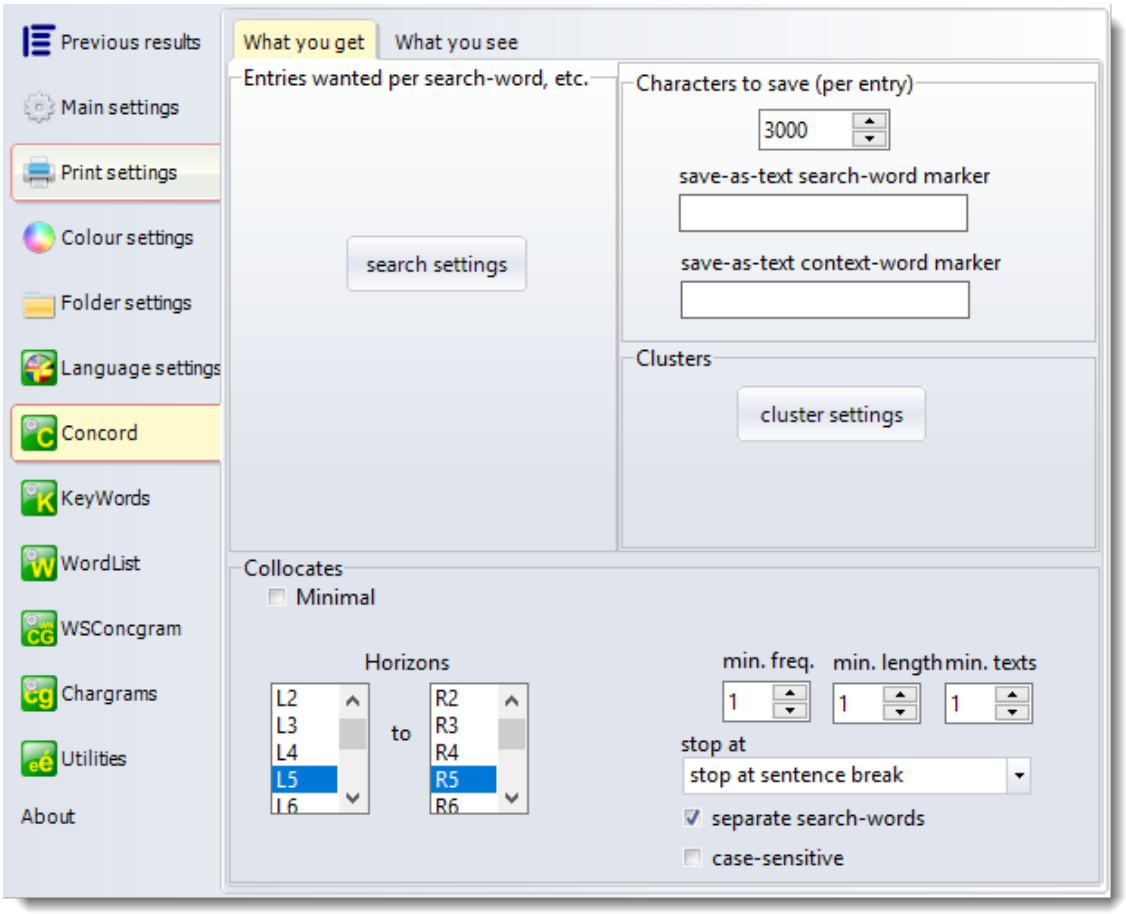
See also: Set column 43

7.17 settings

7.17.1 what you GET



These are found in the main Controller 5 marked *Concord.*
This is because some of the choices -- e.g. collocation horizons 207 -- may affect other Tools.



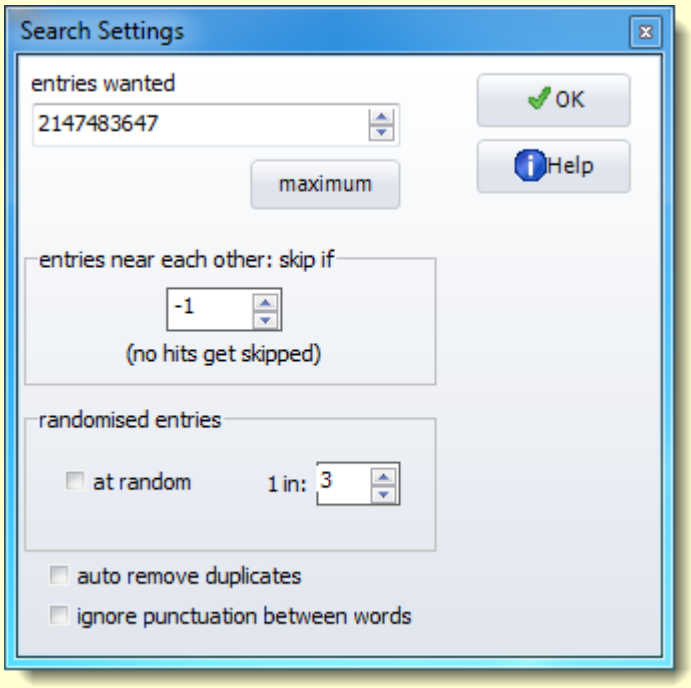
WHAT YOU GET and WHAT YOU SEE

There are 2 tabs for settings affecting *What you get* in the concordance and *What you see* in the display. There is a screenshot at Concord: viewing options 257 showing the options under *What you see*.

WHAT YOU GET

Search Settings

The search settings button lets you choose these settings:



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 (e.g. **book/paperback**), you will get 100 of **book** and 100 of **paperback**.

entries near each other

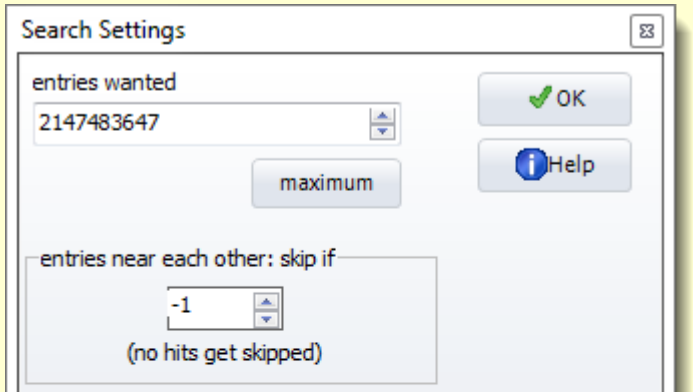
allows you to force Concord to skip hits which are too close to each other. If for example you set this to 0 or 1 and your text contains

... a lovely lovely day

then you will only get the first of these cases if searching for **lovely**. The default here is -1. (If you set it to 0 then you are only allowing one hit within any given word, e.g. a search for ***l*** will only hit each of the **lovely** words once although there are two **l**s in that word.)

randomised entries

This feature allows you to randomise the search. Here Concord goes through the text files and gets the 100 entries by giving each hit a random chance of being selected. To get 100 entries Concord will have to have found around 450-550 hits with the settings shown below. You can set the randomiser anywhere from 1 in 2 to 1 in 1,000. See also: reduce to N 71.



WHAT YOU SEE

The options are explained at Concord: viewing options 257.

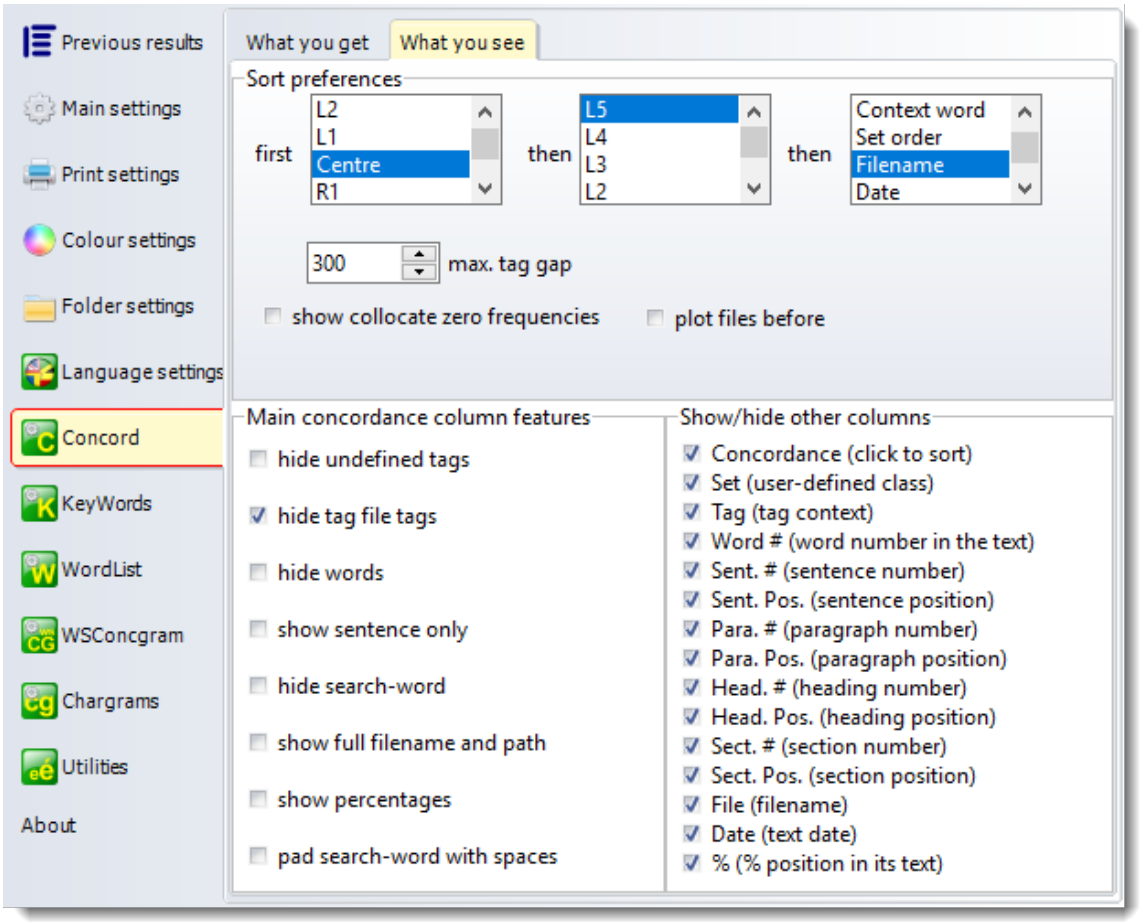
Columns

The *Columns to show/hide* list offers all the standard columns: you may uncheck ones you normally do not wish to see.

See also: Concord Saving and Printing 242, Concord Help Contents 195, Collocation Settings 217.

7.17.2 what you SEE

Access these options in the main Controller, via *Concord* | *What you see*.



Sort preferences

By default, **Concord** will sort a new concordance by the word to the left (L1) then the word to the right (R1), but you can set this to different values if you like. For further details, see Sorting a Concordance 238.

Max. tag gap

When showing tagged text, this controls how many characters around the search-word get coloured 223.

Show collocate zero frequencies

This toggles whether 0 or a blank (the default) is shown if a collocate frequency is zero.

L3	L2	L1	Centre	R1	R2	R3	R4	R5	R6
			27						
						2	1	5	2
1	1			1	2	1	1	1	1
1		1				1	2	1	

or

L3	L2	L1	Centre	R1	R2	R3	R4	R5	R6
0	0	0	27	0	0	0	0	0	0
0	0	0	0	0	0	2	1	5	2
1	1	0	0	1	2	1	1	1	1
1	0	1	0	0	0	1	2	1	0

Concordance View options 196

show raw numbers = show the raw data instead of percentages e.g. for sentence position 263
show full filename and path = sometimes you need to see the whole path but usually the filename alone will suffice.
cut redundant spaces = remove any double spaces
tag string only = show only context within two tag_string_only tags

The sub-set visible shows an opportunity to blank out the search-word, to pad it with a space left & right, to shift the search-word left or right.

See also: Controller 253 *What you get* 253 choices 253, showing nearest tags 223, blanking out 199 the search-word, viewing more context, growing/shrinking concordance lines 199.

7.18 sounds & video

The point of it

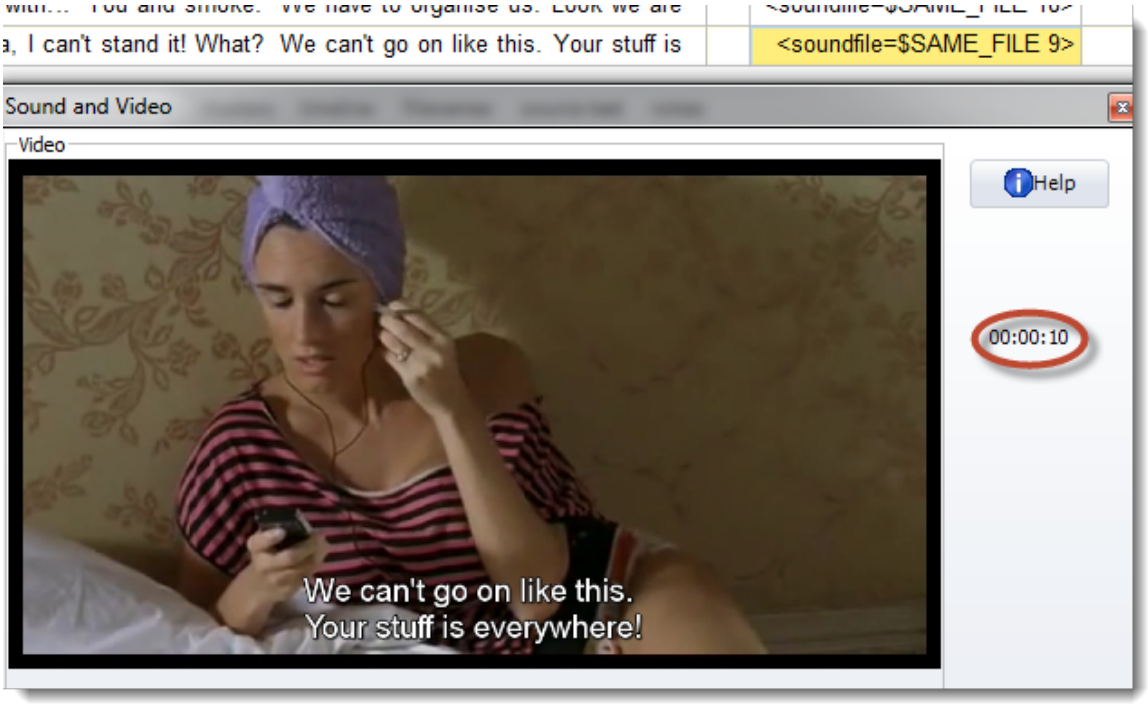
Suppose you do a concordance of "elephant" and want to hear how the word is actually spoken in context. Is the last vowel /?/ or /? /? Does the second vowel sound like "i" or "e" or "u" or a schwa?

How to do it...

If you have defined tags which refer to multimedia files, and if there are any such tags in the "tag-context" of a given concordance line, you can hear or see the source multimedia. The tag will be parsed 162 to identify the file needed, if necessary downloading it from a web address, and then played.

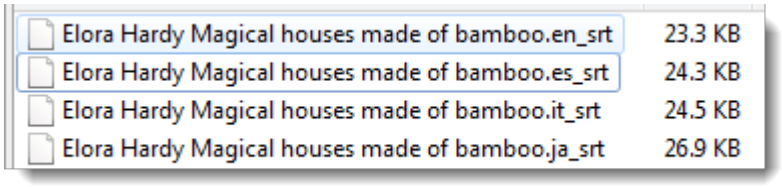
For TED talks 263, use the **srt_video.tag** file supplied with WordSmith.

Video files play if the free VLC Media Player is installed (see http://www.vlcapp.com/). The next screenshot below shows a concordance line with, in the *Tag* column, the mark-up saying that the source text and the video file have the same file-name (except that the latter ends .AVI and the former .TXT). A double-click on the Tag (yellow highlighted cell) brought up the video screen you can see below,

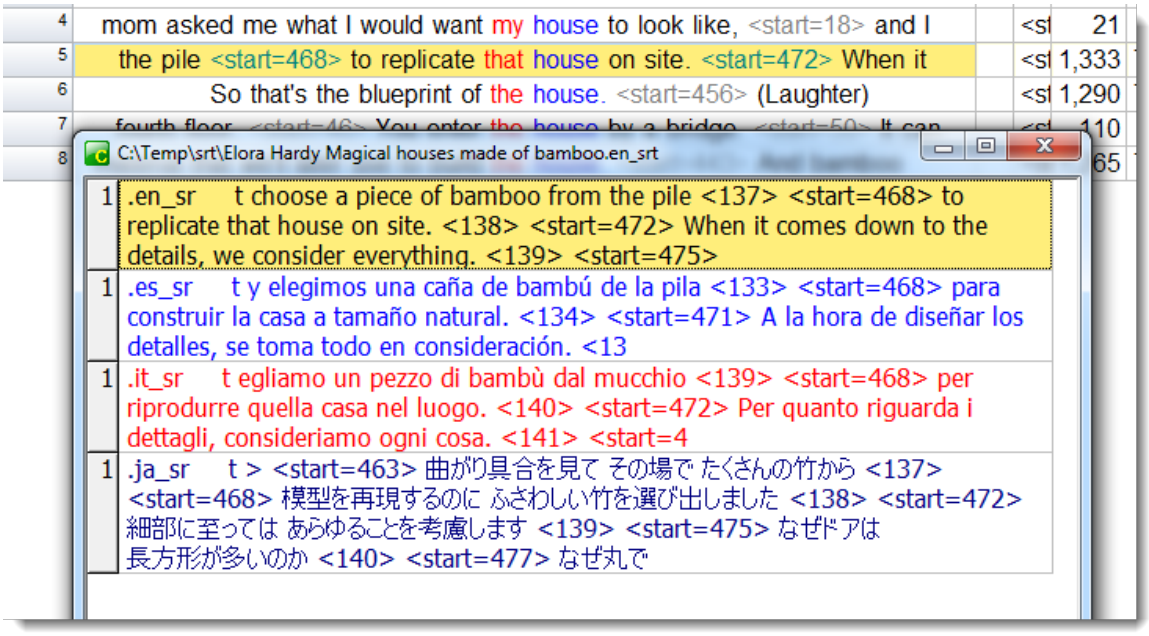


and that has now played to the tenth second, then paused. You can see in the case of this particular video that there is a sub-title with the same words that are in the concordance above (though there is no guarantee you will see sub-titles for all videos).

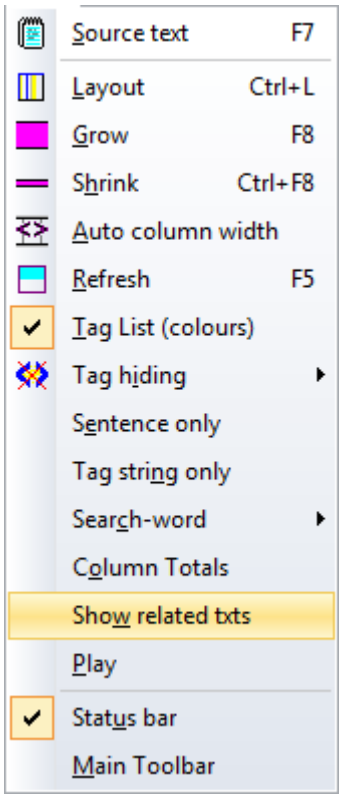
If you build up a collection of TED talks like these where the same video in English has transcripts in several languages,



you can get to see the different translations:



by choosing *View* | *Show related txts* in the menu.

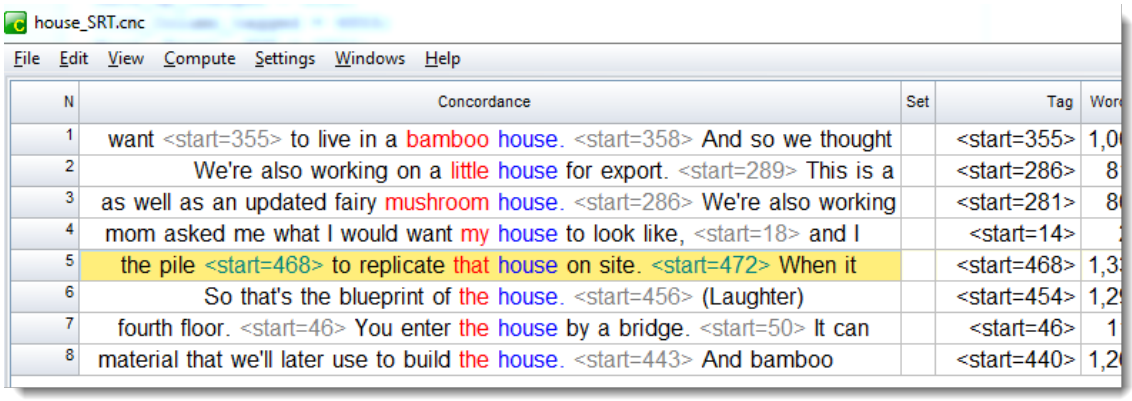


Note: If you want to see the whole text of a given line, double-click it and the relevant translation text file will open in Notepad.

You can also play video files in Aligner[391] and can do parallel concordancing[396] (where you see a word in one language and its translations in others).

See also: Multi-media Tag syntax[162], Playing the video[261], Obtaining Sound and Video files[263], Handling Tags[158], Making a Tag File[172], Showing Nearest Tags in Concord[223], Tag Concordancing[221], Types of Tag[159], Viewing the Tags[401], Using Tags as Text Selectors[167], Tags in WordList[367]

7.18.1 playing the video



In this screenshot we see a concordance where there is a tag inserted periodically in the text file. To play the media file, select a line, choose *File* | *Play media file*, or double-click the *Tag* column.

Here we see a video being shown in Concord after a concordance tag column was double-clicked:

The timing visible shows we have reached the 108th second of the SRT transcript:

```
<start=98>
To me, that did not add up.

<start=102>
It was too unreal.

<start=106>
So when I was 11, I became ill.

<start=108>
I fell into depression,

<start=111>
I stopped talking,
```

+/-

The +/- box lets you adjust the timing of the transcript in the video player, if the SRT timings are wrong by a fixed amount, e.g. if 108 is not where she says she fell into depression. You might for example set this to **-5** to go 5 seconds earlier than the transcript says every time you double-click to play a concordance extract.

Buttons

The *Slower* and *Faster* buttons adjust the rate of speech. *<1 sec* and *>1 sec* let you go back or forward a second after pausing.

See also: Multi-media Tag syntax 162, Sound and Video 258, Obtaining Sound and Video files 263.

7.18.2 obtaining sound and video files

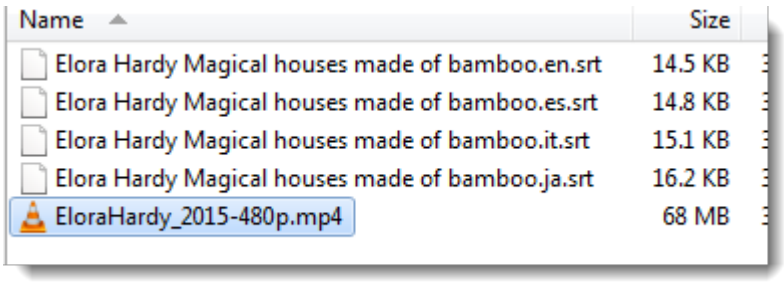
Sources of sound and video files

WordSmith does not provide or include corpora. However, there are specialised corpora such as NECTE, MICASE, ICE and then there are publicly available sources such as the TED Talks. You are expected to respect copyright provisions in all cases.

There is a lot of useful advice at

TED Open Translation Project where you will find transcripts.

These text files in English (.en), Spanish (.es), Italian (.it) and Japanese (.ja) were downloaded from there and later converted using the Text Converter 502

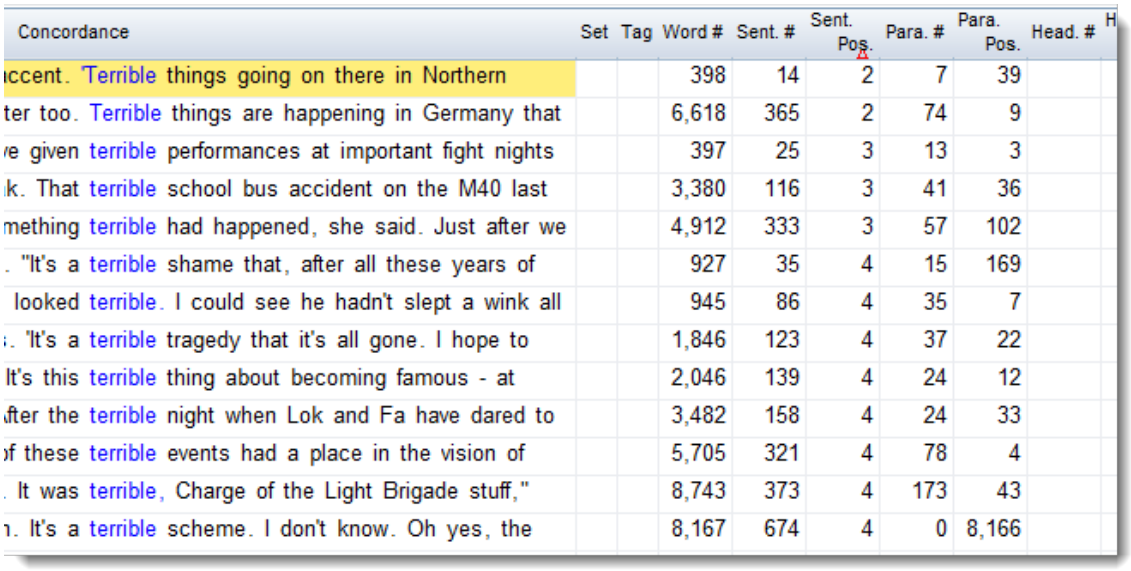


If you wish to use a transcript and sound file format which is incompatible with the syntax 162 described here, please contact 532 us.

7.19 text segments in Concord

A concordance line brings with it information about which segment of the text it was found in.

In the screenshot below, a concordance was carried out; the listing has been sorted by Sentence Position. 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.



Concordance	Set	Tag	Word #	Sent. #	Sent. Pos.	Para. #	Para. Pos.	Head. #	H
ıccent. Terrible things going on there in Northern			398	14	2	7	39		
ter too. Terrible things are happening in Germany that			6,618	365	2	74	9		
ve given terrible performances at important fight nights			397	25	3	13	3		
k. That terrible school bus accident on the M40 last			3,380	116	3	41	36		
mething terrible had happened, she said. Just after we			4,912	333	3	57	102		
. "It's a terrible shame that, after all these years of			927	35	4	15	169		
looked terrible. I could see he hadn't slept a wink all			945	86	4	35	7		
. 'It's a terrible tragedy that it's all gone. I hope to			1,846	123	4	37	22		
It's this terrible thing about becoming famous - at			2,046	139	4	24	12		
ıfter the terrible night when Lok and Fa have dared to			3,482	158	4	24	33		
of these terrible events had a place in the vision of			5,705	321	4	78	4		
. It was terrible, Charge of the Light Brigade stuff,"			8,743	373	4	173	43		
ı. It's a terrible scheme. I don't know. Oh yes, the			8,167	674	4	0	8,166		

(With the settings when this concordance was computed, no headings were identified because there was no suitable mark-up in the plain text, so the heading position and number for each line is blank.)

You can toggle the numbers between raw numbers and percentages 257. You can switch the numbers on or off 256 as preferred.

See also: Start and end of text segments 164.

KeyWords

8 KeyWords

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 272.

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 292 between them can be plotted 296, made into a database 275, and grouped according to their associates 279.

Online step-by-step guide showing how

8.1 index

Explanations

What is the Keywords program and what's it for? 266
How Key Words are Calculated 287
2-Word list Analysis 267
Key words display 302
Key words plot 296
Key words plot display 294
Plot-Links 292
Batch Analyses 30
Database of Key Key-Words 275
- Associates 279
- Follow-up 282
- Limitations 544

Settings and Procedures

Calling up a Concordance 69
Choose Word Lists 269
Colours 114

Database 276
Folders 539
Fonts 118
Keyboard Shortcuts 546
Printing 113
Re-sorting 296
Exiting 91

Tips

KeyWords advice 286
Window management 151

Definitions

General Definitions 533
Key-ness 273
Key key-word 278
Associate 281

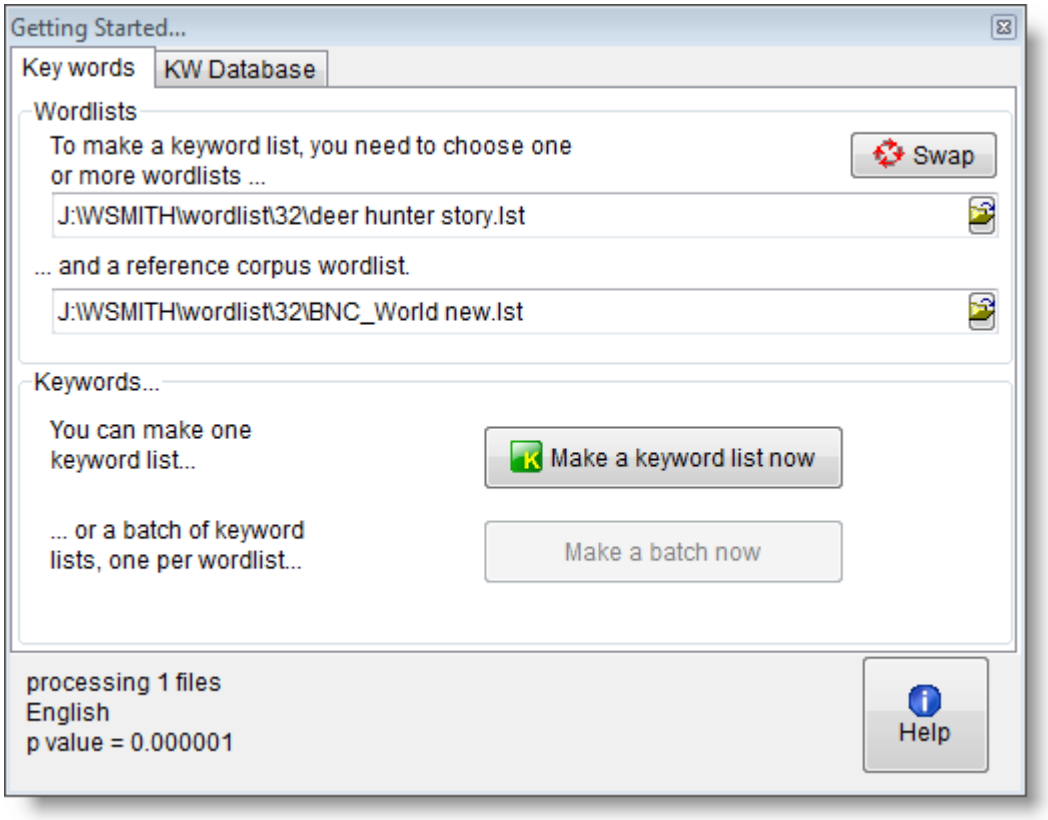
See also : WordSmith Main Index 2

8.2 ordinary two word-list analysis

The usual kind of **KeyWords** analysis. It compares the one text file (or corpus) you're chiefly interested in, with a reference corpus 555 often based on a lot of text or else a comparable one. In the screenshot below we are interested in the key words of **deer hunter story** and we're using **BNC** as the reference corpus to compare with.

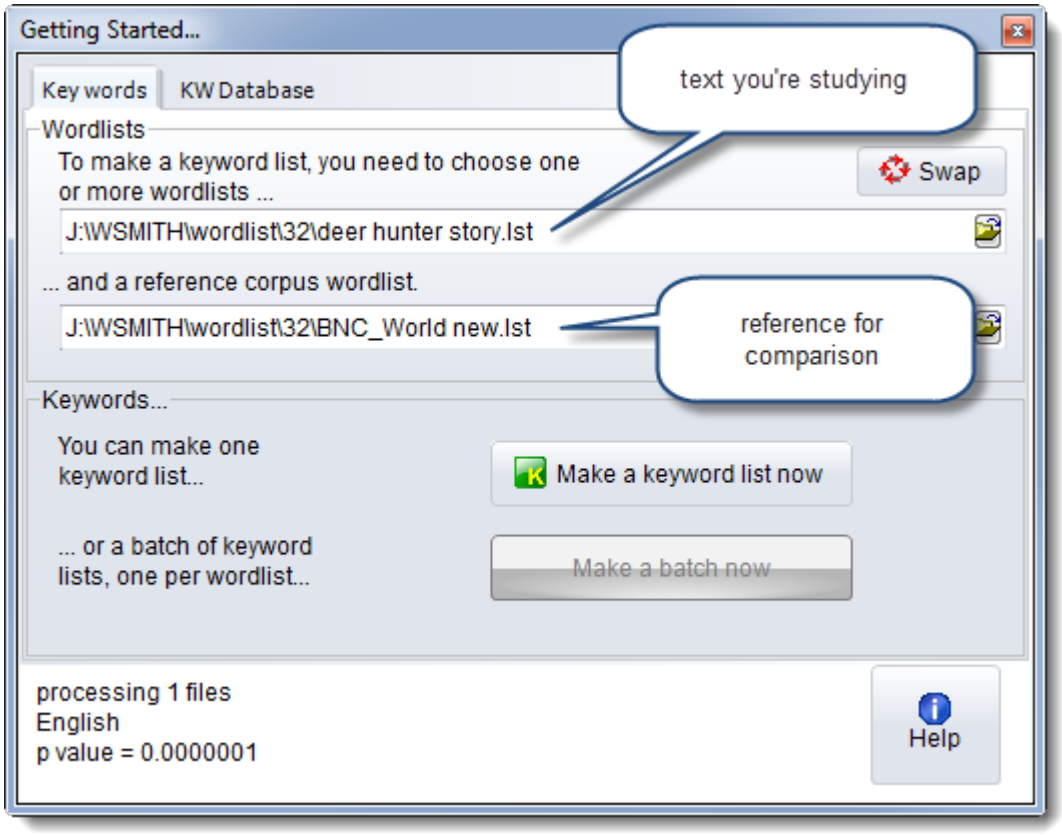
Choose Word Lists

In the dialogue box you will choose 2 files. The text file in the box above and the reference corpus file in the box below.



See also How Key Words are Calculated 287, KeyWords Settings 297

8.3 choosing files



Current Text word list

In the upper box, choose a word list file.

To choose more than 1 word list file, press Control as you click to select non-adjacent lists, or Shift to select a range.

This box determines which word-list(s) you're going to find the key words of.

Reference Corpus word list

The the box below, you choose your Reference Corpus 555 List. (This can be set permanently in the main Controller Settings).

No word lists visible

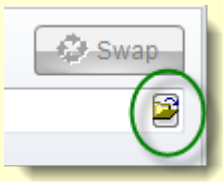
If you can't see any word lists in the displays, either change folders until you can, or go back to the WordList tool and make up at least 2 word lists: this procedure requires at least two before it can make a comparison.

Swap

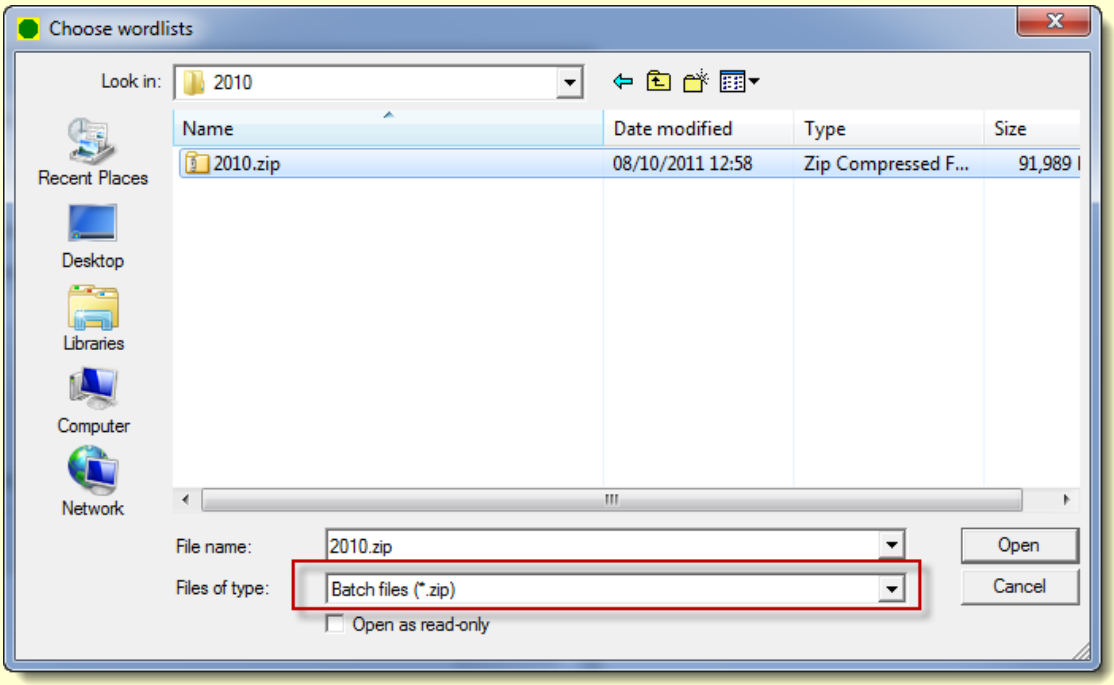
The text you're studying must be at the top. If you get them wrong, exchange them.

Advanced: working with a batch file

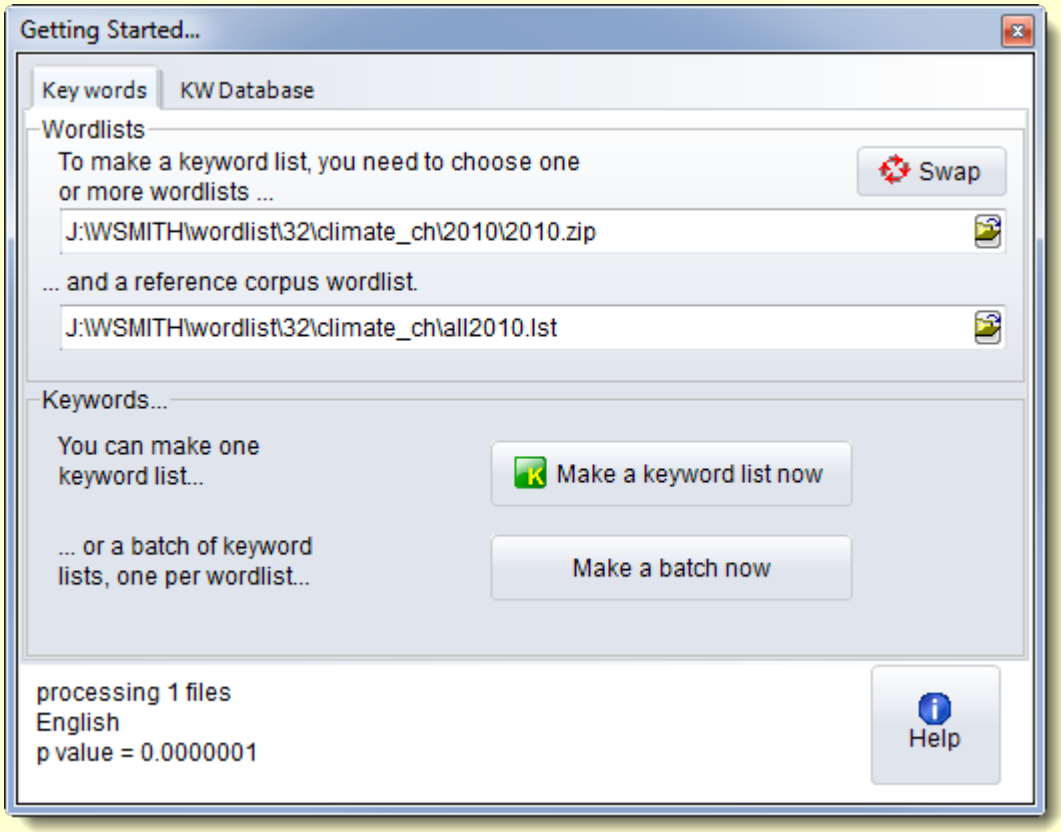
Click the browse button:



and choose the batch **`.zip`** file



and we are ready to make a batch: that one 2010.zip contains many thousands of word lists.



8.4 example of key words

Here is an example from the play Othello.

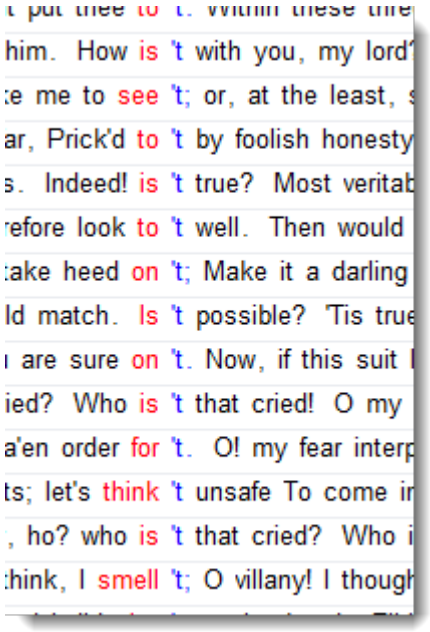
These "key" words are not the most frequent words (which are those like *the*) but the words which are most unusually frequent in the Shakespeare play. Key words usually give a reasonably good clue to what the text is about.

J:\WSMITH\KEYWORDS\32\othello, the moor of venice.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	BIC	Log_L	Log_R	P
1	CASSIO	113	0.43	1	133	0.02	450.10	463.78	4.80	0.0000000000
2	IAGO	60	0.23	1	88	0.01	213.99	227.67	4.48	0.0000000000
3	MOOR	56	0.21	1	78		202.94	216.63	4.55	0.0000000000
4	DESDEMONA	40	0.15	1	59		137.77	151.46	4.47	0.0000000000
5	T	71	0.27	1	450	0.05	98.15	111.84	2.37	0.0000000000
6	HANDKERCHIEF	28	0.11	1	36		97.80	111.48	4.67	0.0000000000
7	LIEUTENANT	29	0.11	1	51		88.68	102.37	4.22	0.0000000000
8	RODERIGO	27	0.10	1	47		82.03	95.72	4.23	0.0000000000
9	CYPRUS	23	0.09	1	30		77.42	91.11	4.65	0.0000000000
10	OTHELLO	24	0.09	1	45		68.75	82.44	4.13	0.0000000000
11	WILLOW	18	0.07	1	25		56.02	69.70	4.56	0.0000000000
12	HONEST	42	0.16	1	295	0.03	46.17	59.85	2.22	0.0000000000

Here the KWs show characters, themes. The word **T** is an abbreviation of **it** :



Per 100, 1000 or million?

In advanced settings you can display the % columns 131 in different formats:

Key word	Freq.	per 1000 words	Texts	RC. Freq.	per 1000 words	BIC
ROMEO	115	4.77	1	130	0.15	481.57
TYBALT	47	1.95	1	52	0.06	190.11

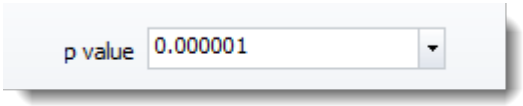
See also: word-lists with tags as prefix 374.

8.5 keyness

8.5.1 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.

Setting the p value



In the case of key item analyses, where the notion of risk combines with selectivity and strength of relationship, you may often wish to set a comparatively stringent ***p*** value threshold such as 0.000001 (one in 1 million) (1E-6 in scientific notation). A p value alone tells you nothing about whether the relationship is weak or strong, however and if you are going to be guided by BIC scores 288 you are recommended to leave the p value at 0.1.

Alternative to p value

So it is better to set a suitable BIC score 288 threshold such as 2.5 or greater, measuring the strength of the keyness and not just its significance.

You can also set a low "maximum wanted" number in the main Controller 5, under *KeyWords Settings* 297.

See also: Log ratio 288, BIC score 288, Log Likelihood 288, Definitions 533, choosing a reference corpus 274

8.5.2 keyness 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[555] is such that the statistical probability as computed by an appropriate procedure[287] is smaller than or equal to a p value[273] specified by the user
c) in addition, the strength of keyness must be at least as great as the minimum log ratio[288] set by the user.

positive and negative keyness

A word which is *positively* key occurs more often than would be expected by chance in comparison with the reference corpus.

A word which is *negatively* key occurs less often than would be expected by chance in comparison with the reference corpus.

typical key words

KeyWords will usually throw up 3 kinds of words as "key".

First, there will be proper nouns. Proper nouns are often key in texts, though a text about racing could wrongly identify as key, names of horses which are quite incidental to the story. This can be avoided by specifying a higher Minimum Frequency.

Second, there are key words that human beings would recognise. The program is quite good at finding these, and they give a good indication of the text's "aboutness". (All the same, the program does not group synonyms, and a word which only occurs once in a text may sometimes be "key" for a human being. And KeyWords will not identify key phrases unless you are comparing word-lists based on word clusters[556].)

Third, there are high-frequency words like **because** or **shall** or **already**. These would not usually be identified by the reader as key. They may be key indicators more of style than of "aboutness". But the fact that KeyWords identifies such words should prompt you to go back to the text, perhaps with Concord (just choose *Compute* | *Concordance*), to investigate why such words have cropped up with unusual frequencies.

See also: How Key Words are Calculated[287], Definition of Key Key-Word[278], Definitions[533], KeyWords Settings[297]

8.5.3 thinking about keyness

Choosing a reference corpus

In general the choice does not make a lot of difference if you have a fairly small p value (such as 0.000001). But it may help to think using this analogy.

Different reference corpora may give different results. Suppose you have a method for comparing objects and you take a particular apple out of your kitchen to compare using it

A) with a lot of apples in the green-grocer's shop

B) with all the fruit in the green-grocer's shop

C) with a mixture of objects (cars, carpet, notebooks, fruit, elephants etc.)

With A) you will get to see the individual characteristics, e.g. perhaps your apple is rather sweeter than most apples. (But you won't see its "apple-ness" because both your apple and all the others in your reference corpus are all apples.)

With B) you will see "appleness" (your apple, like all apples but unlike bananas or pineapples, is rather round and has a very thin skin) but might not see that your apple is rather sweet and you won't get at its "fruitiness".

With C) you will get at the apple's fruity qualities: it is much sweeter and easier to bite into than cars and notebooks etc.

Keyness scores

Is there an important difference between a key word with a keyness of 50 and another of 500?

Suppose you process a text about a farmer growing 3 crops (wheat, oats and chick-peas) and suffering from 3 problems (rain, wind, drought). If each of these crops is equally important in the text, and each of the 3 problems takes one paragraph each to explain, the human reader may decide that all three crops are equally key and all three problems equally key. But in English these three crop-terms and weather-terms vary enormously in frequency (chick-peas and drought least frequent). WordSmith's KW analysis will necessarily give a higher keyness value to the rarer words. So it is generally unsafe to rely on the order of KWs in a KW list.

8.6 KeyWords database

(default file extension .KDB)

The point of it…

The point of this database is that it will allow you to study the key-words which recur often over a number of files.

For example, if you have 500 business reports, each one will have its own key words. These will probably be of two main kinds. There will be key-words which are key in one text but are not generally key (names of the firms and words relating to what they individually produce); and other, more general words (like **consultant, profit, employee**) which are typical of business documentation generally. Or you may find that **I, you, should** etc. come to the top if your text files are ones which are much more interactive than the reference corpus 555 texts.

By making up a database, you can sort these out. The ones at the top of the list, when you view them, may be those which are most typical of the genre in some way. We might call the ones at the top "key-key words" and the list is at first ordered in terms of "key key-ness", but those at the bottom will only be key in a few text files. You can of course toggle it into alphabetical order and back again.

You can set a minimum number of files that each word must have been found to be key in, using *KeyWords Settings* | *Database* 300.

When viewing a database you will be able to investigate the associates 279 of the key key-words.

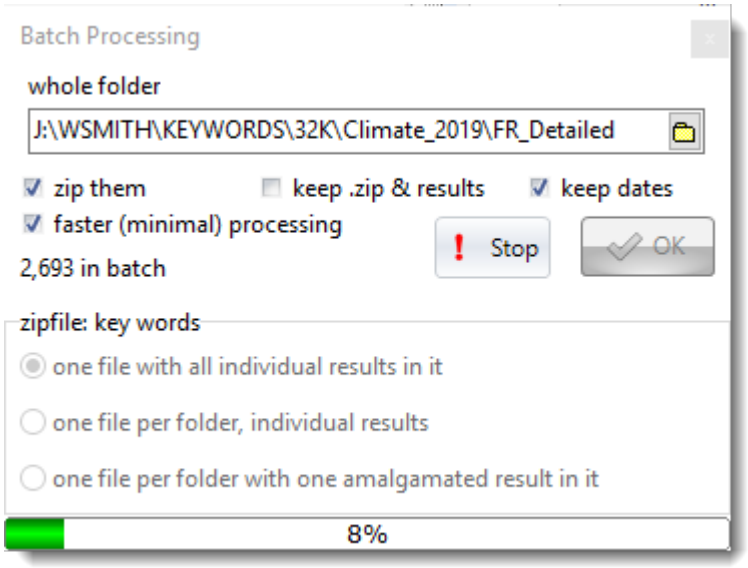
Under Statistics, you will also be able to see details of the key words files which comprise the database (file name and number of key words per file), together with overall statistics on the number of different types and the tokens (the total of all the key-words in the whole database including repeats).

See also : Creating a database 276, Definition of key key-word 278

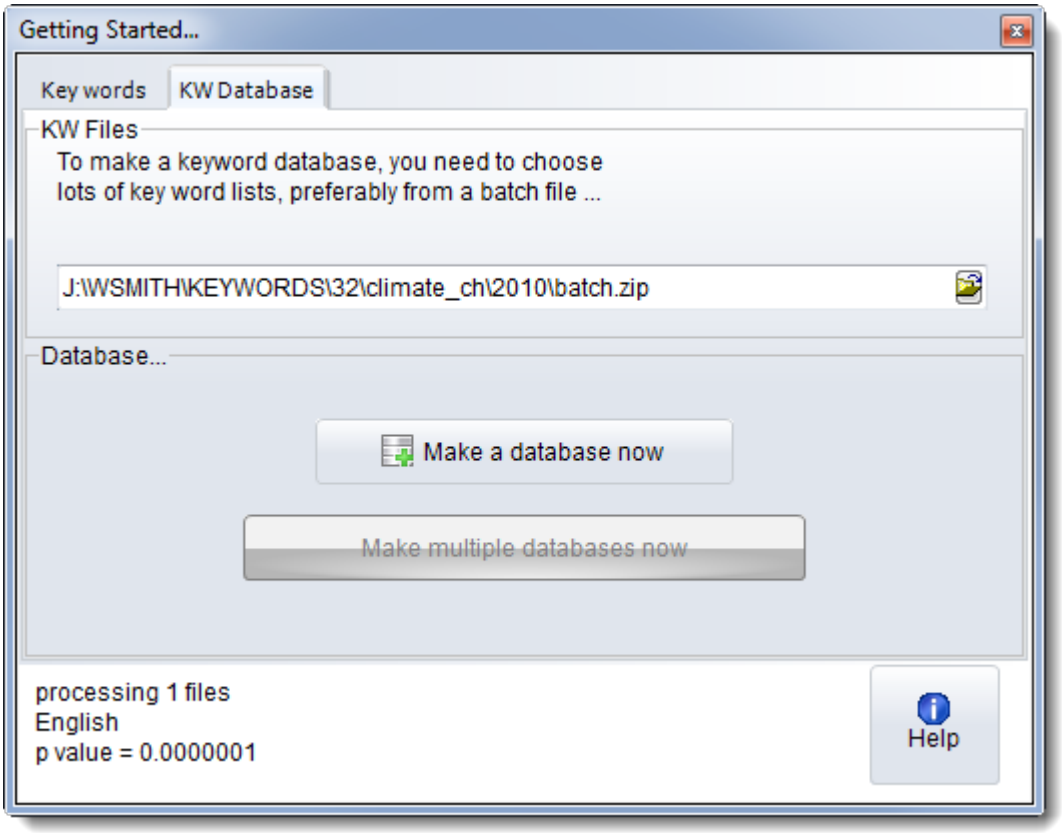
8.6.1 creating a database

To build a key words database, you will need a set of key word lists. For a decent sized database, it is preferable to build it like this:

1. Make a batch 30 of word lists.
2. Use this batch off word lists to make a batch 30 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* | *KW Database*.



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

Note that making a database means that only positive 287 key words will be retained.

The process takes a while as it first reads in each of the components of your .zip file(s), then it computes all the associates 279.

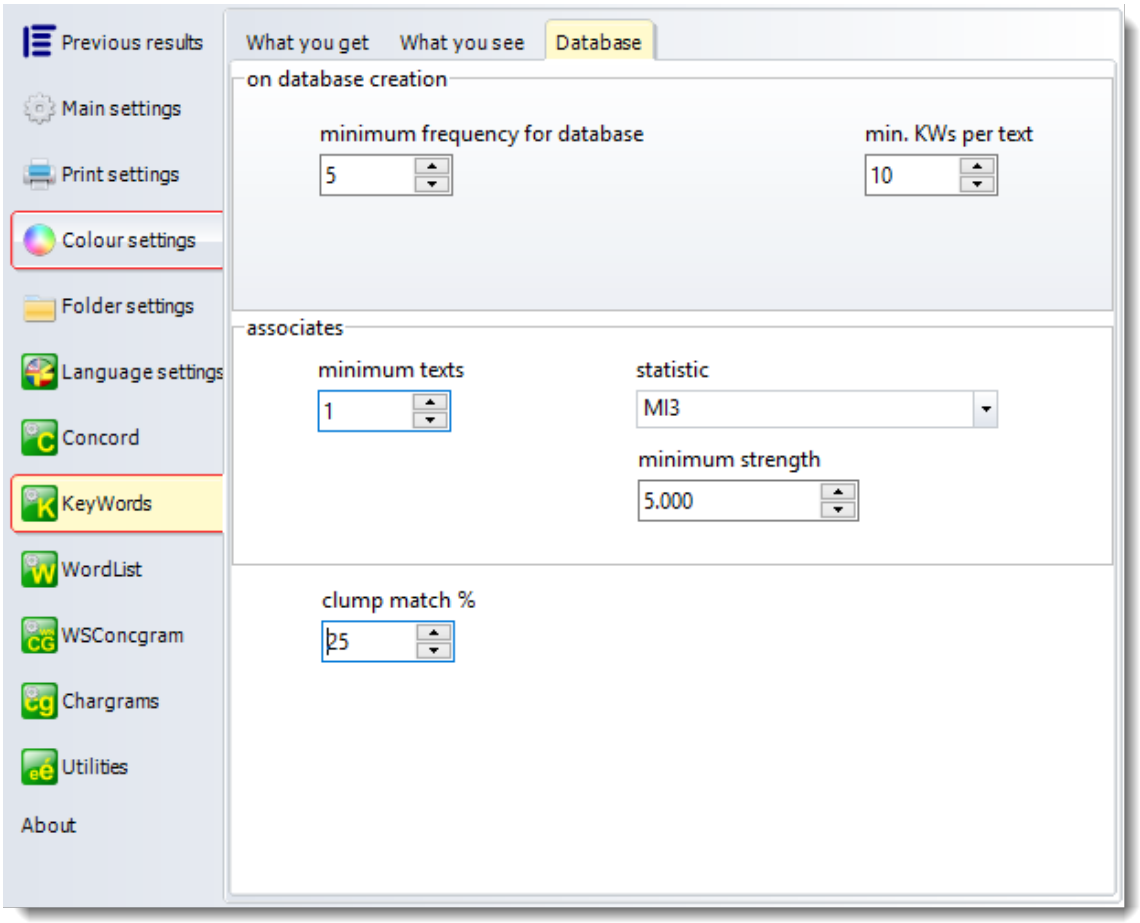
In the Controller KeyWords settings 297 you can make other choices:

minimum frequency for database

If you set this to 5 you will only use for the database any KWs which appear in 5 or more texts

min. KWs per text

If this is set to 10, any KW results files which ended up with very few positive KWs will be ignored.



See also: associates 279.

8.6.2 key key-word definition

A "key key-word" is one which is "key" in more than one of a number of related texts. The more texts it is "key" in, the more "key key" it is. This will depend a lot on the topic homogeneity of the corpus being investigated. In a corpus of City news texts, items like *bank*, *profit*, *companies* are key key-words, while *computer* will not be, though *computer* might be a key word in a few City news stories about IBM or Microsoft share dealings.

Requirements

To discover "key key words" you need a lot of text files (say 500 or more), ideally fairly related in their topics, which you make word-lists of (it's much faster doing that in a batch), and then you have to compute key word-lists of each of those, all of which go into a database. It is all explained under creating a keywords database 276.

See also: How Key Words are Calculated 287, Definition of Key Word 273, Creating a Database 276, Definitions 533

8.6.3 associates

"Associates" is the name given to key-words associated with a key key-word[275].

The point of it…

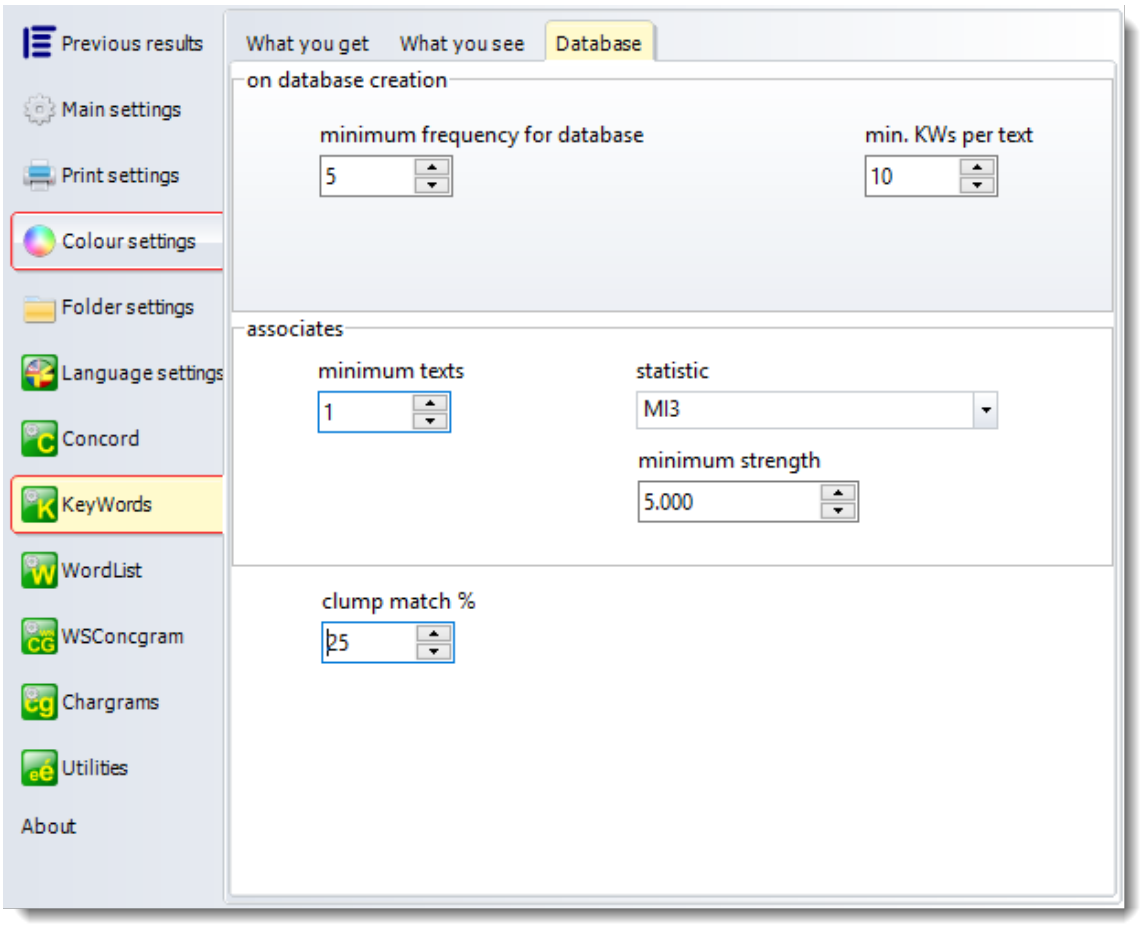
The idea is to identify words which are commonly associated with a key key-word, because they are key words in the same texts as the key key-word is. An example will help.

Suppose the word *wine* is a key key-word in a set of texts, such as the weekend sections of newspaper articles. Some of these articles discuss different wines and their flavours, others concern cooking and refer to using wine in stews or sauces, others discuss the prices of wine in a context of agriculture and diseases affecting vineyards. In this case, the associates of *wine* would be items like *Chardonnay, Chile, sauce, fruit, infected, soil*, etc.

The listing shows associates in order of frequency. A menu option allows you to re-sort them.

Settings

You can set a minimum number of text files for the association procedure, in the database settings[297]:



Minimum texts

These screenshot settings would only process those key-key-words which appear in at least 1 text file.

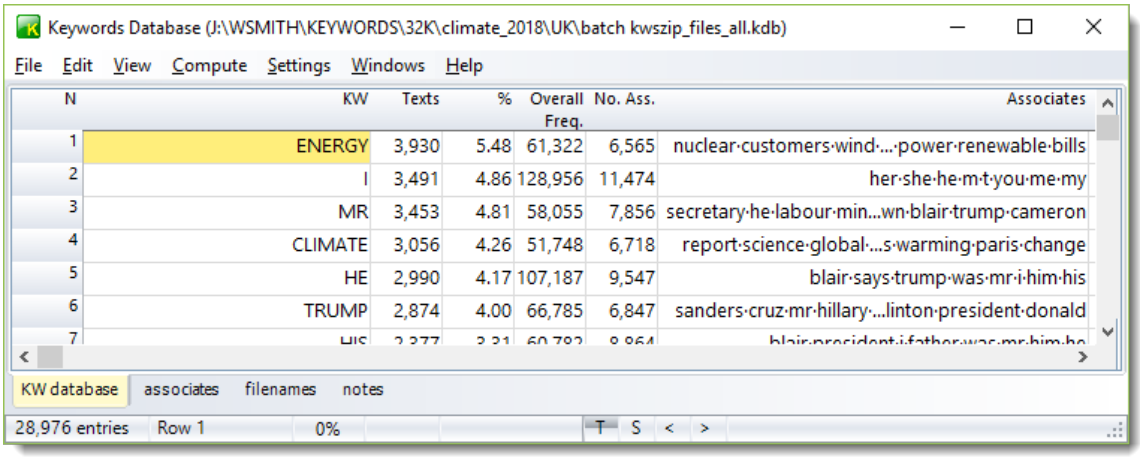
Statistic

Choose the mutual information statistic 342 you prefer, apart from Z score which uses a span (here we're using the whole text).

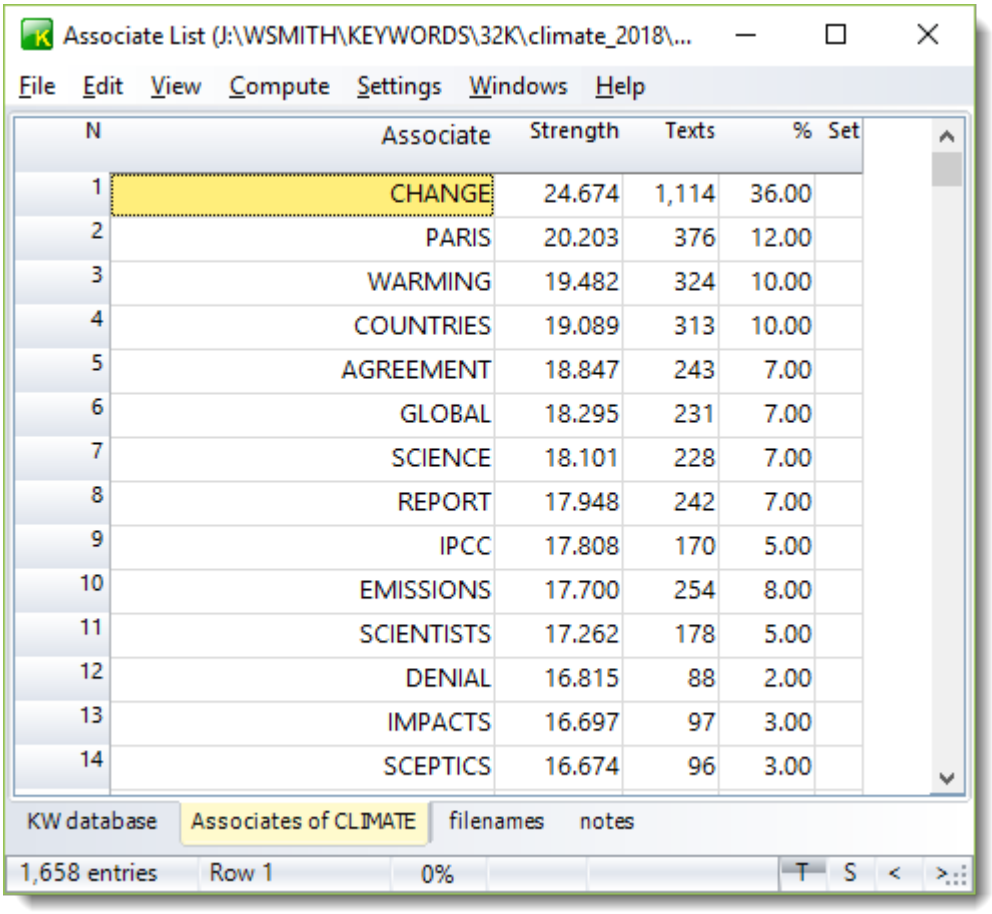
Minimum strength

This will only show associates which reach at least the strength in the statistic set here, e.g. 5.000.

This screenshot shows the most frequent associates in the right-hand column of the main keywords data base window. **`Energy`** goes with **`bills`**, with **`renewable`**, etc. That is, in texts where **`energy`** is key, **`bills`** tends to be key too.



To see the detailed associates, double-click your chosen term in the KW column or the Associates column:



N	Associate	Strength	Texts	%	Set
1	CHANGE	24.674	1,114	36.00	
2	PARIS	20.203	376	12.00	
3	WARMING	19.482	324	10.00	
4	COUNTRIES	19.089	313	10.00	
5	AGREEMENT	18.847	243	7.00	
6	GLOBAL	18.295	231	7.00	
7	SCIENCE	18.101	228	7.00	
8	REPORT	17.948	242	7.00	
9	IPCC	17.808	170	5.00	
10	EMISSIONS	17.700	254	8.00	
11	SCIENTISTS	17.262	178	5.00	
12	DENIAL	16.815	88	2.00	
13	IMPACTS	16.697	97	3.00	
14	SCEPTICS	16.674	96	3.00	

See also: definition of associate 281, related clusters 285.

8.6.3.1 associate definition

An "associate" of key-word X is another key-word (Y) which co-occurs with X in a number of texts. It may or may not co-occur in proximity to key-word X. (A *collocate* would have to occur within a given distance of it, whereas an associate is "associated" by being key in the same text.)

For example, in a key-word database of *Guardian* newspaper text, *wine* was found to be a key word in 25 out of 299 stories from the Saturday "tabloid" page, thus a key key word 278 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 273, Associates 279, Definitions 533, Mutual Information 342

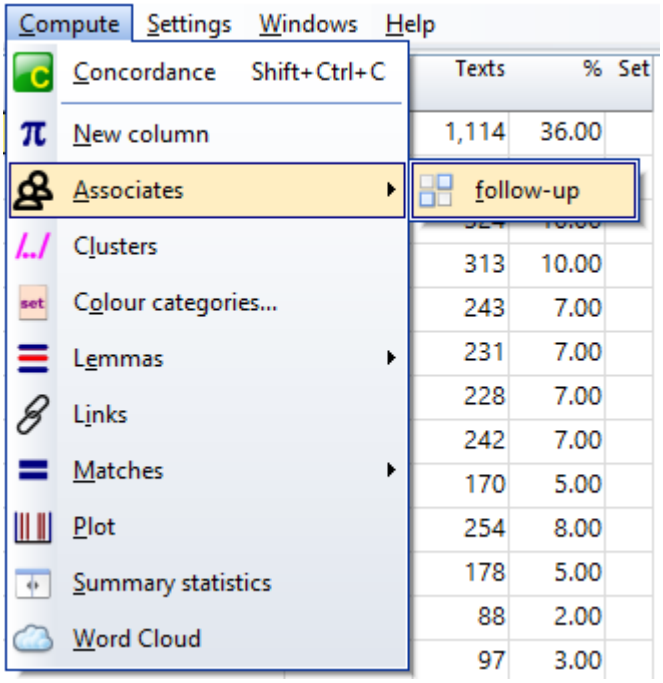
8.6.3.2 follow-up

The point of it...

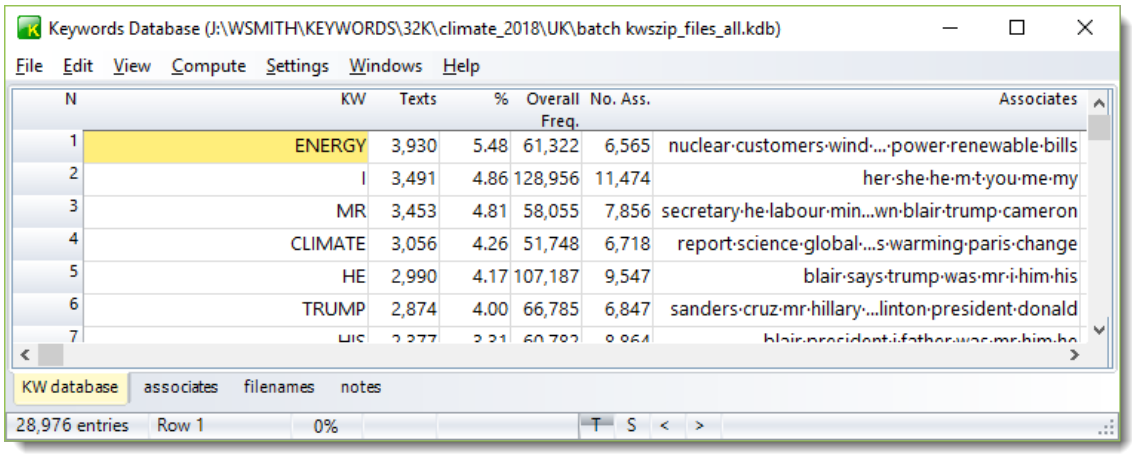
The idea is to follow up a set of associates by identifying the texts which are most closely related to the topic it suggests.

How to do it

in a list of associates, choose the *follow-up* menu item.



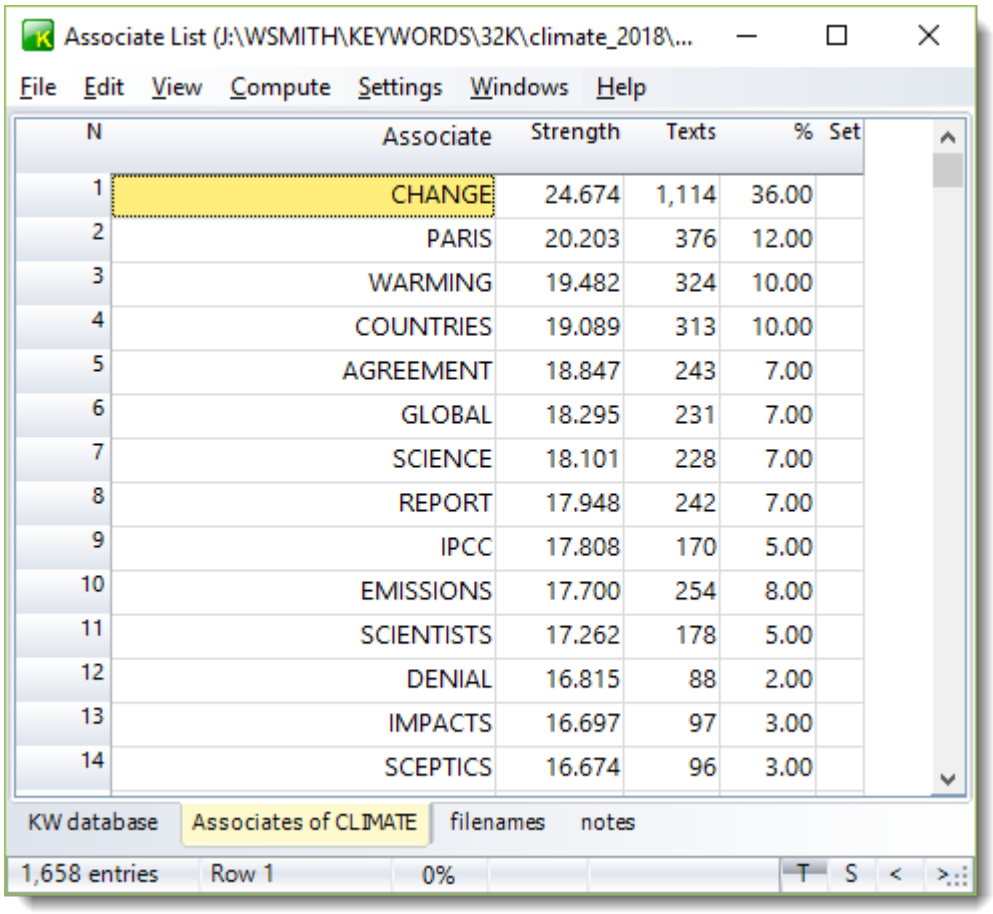
Example



Here, a study on climate change texts has **`climate`** as one of its key key words. Among the 75,000 source texts, some mention climate change only incidentally and I wanted to restrict the

corpus by about half, separating out such texts. (For another purpose one might want to study only those texts of course.)

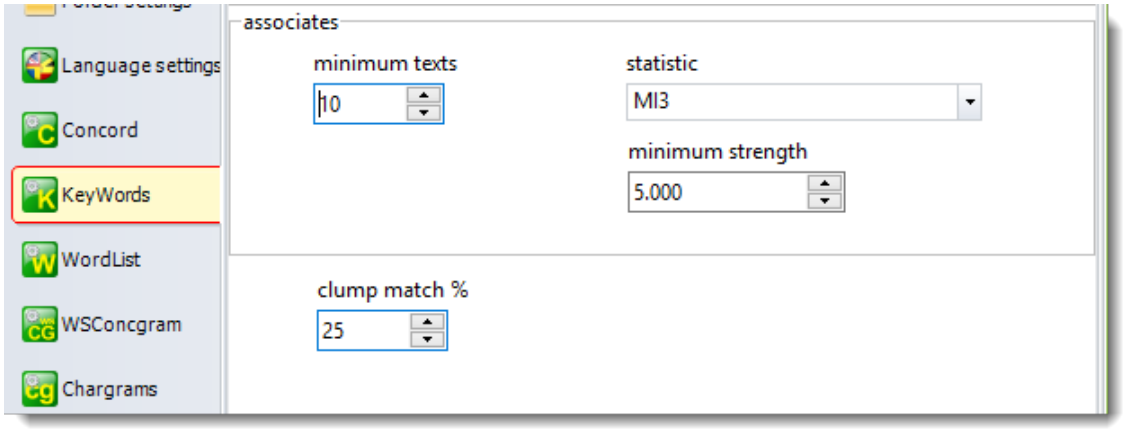
A first step is to identify the associates of **climate**.



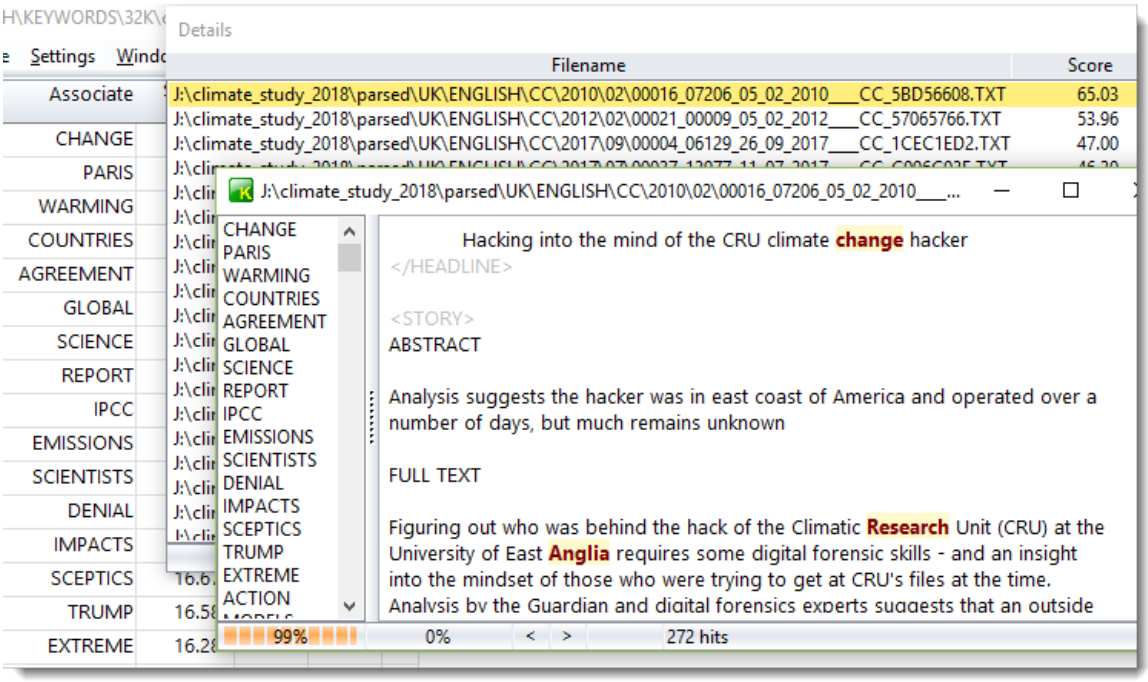
N	Associate	Strength	Texts	%	Set
1	CHANGE	24.674	1,114	36.00	
2	PARIS	20.203	376	12.00	
3	WARMING	19.482	324	10.00	
4	COUNTRIES	19.089	313	10.00	
5	AGREEMENT	18.847	243	7.00	
6	GLOBAL	18.295	231	7.00	
7	SCIENCE	18.101	228	7.00	
8	REPORT	17.948	242	7.00	
9	IPCC	17.808	170	5.00	
10	EMISSIONS	17.700	254	8.00	
11	SCIENTISTS	17.262	178	5.00	
12	DENIAL	16.815	88	2.00	
13	IMPACTS	16.697	97	3.00	
14	SCEPTICS	16.674	96	3.00	

Now in the main controller key words database settings, choose a clump match percentage. The higher this is, the smaller the list of files we get will be and the more selective the criterion.

The output will now be a list of words which you could edit and use for the relevance check 429 (Corpus Checker tool).



Choose *Associates* | *follow-up* in the menu, and you'll get a list of files. If you double-click one, you get to read it. As you see, the text does seem to be one which is definitely about climate change.

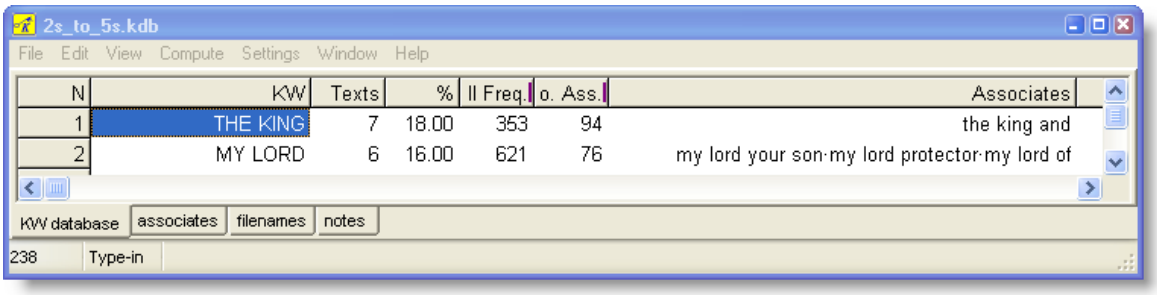


The list of terms at the left is the terms which were determined most relevant in the follow-up. Right-click to get these terms highlighted.

You can save the list as a file (right-click to see the menu). This could be input to the Corpus Checker sampler 434 so you can evaluate your results.

8.6.4 keywords database related clusters

The idea is to be able to find any overlapping clusters in a key word database, e.g. where MY LORD is related to MY LORD YOUR SON.



To achieve this, choose *Compute* | *Clusters*. To clear the view, *Compute* | *Associates*.

See also: associates 279

8.6.5 export data

If you have a keywords database open, you can export its data using *File* | *Export* data...

This will export a list of the text files, followed by a list of the words in the database

```
        text  (475) \\DISKSTATIONTWO\Mike\text\480texts\TA000255.TXT
        text  (476) \\DISKSTATIONTWO\Mike\text\480texts\TA000276.TXT
        text  (477) \\DISKSTATIONTWO\Mike\text\480texts\TA000291.TXT
        text  (478) \\DISKSTATIONTWO\Mike\text\480texts\TA000307.TXT

Keywords Database
12,220 words
        word 1 = #
        word 2 = A
        word 3 = A'S
        word 4 = ABABA
        word 5 = ABALKIN
        word 6 = ABATEMENT
```

then the details of each word with the number of each file it is found in and the frequency; thus in these data,

```
        word 12,218 = ZOOLOGISTS
        word 12,219 = ZULU
        word 12,220 = ZURBRIGGEN

12,220 words: text file (frequency)
#
                1 (166) 2 (163) 4 (206) 7 (224) 11 (342)        13 (251)
A
                17 (218)        112 (180)       174 (315)       318 (82)
A'S
                40 (3)  162 (3) 244 (3)
ABABA
                9 (7)
ABALKIN
                100 (6)
```

A'S was found in 3 text files (file 40, 162 and 244 listed above) and in each of those text files it occurred three times. The list is formatted with tab characters between each pair so you could import the data into Excel if required.

8.7 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 230.

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 555.

3. Compare apples with pears, or, better still, Coxes with Granny Smiths. So choose your reference corpus in some principled way 274. The computer is not intelligent and will try to do whatever comparisons you ask it to, so it's up to you to use human intelligence and avoid comparing apples with phone boxes!

If it didn't work...

For the procedure to work, a number of conditions must be right: the language 77 defined for each word list must be the same (that is, Mexican Spanish and Iberian Spanish count as the same but Iberian Spanish and Brazilian Portuguese count as different so could not be compared in this process); each word list must have been sorted alphabetically 372 in ascending order before the comparison is made. (The program tries to ensure this, automatically.) Also, any prefixes 374 or suffixes must match.

8.8 KeyWords: calculation

The "key words" are calculated by comparing the frequency of each word in the word-list of the text you're interested in (study corpus) with the frequency of the same word in another word-list (comparison corpus 555). All words which appear in the smaller list are considered, unless they are in a stop list 141.

If **the** occurs say, 5% of the time in the study corpus and 6% of the time in the comparison 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 comparison corpus (unless your comparisonr corpus only concerns spiders!)

To compute the "keyness" of an item, the program therefore computes

its frequency in the small word-list
the number of running words 349 in the small word-list
its frequency in the other corpus
the number of running words 349 in the comparison corpus

and cross-tabulates these.

A word will get into the listing here if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the comparison word-list.

Unusually infrequent key-words are called "negative key-words" and appear at the very end of your listing, in a different colour. Note that negative key-words will be omitted automatically from a keywords database 275 and a plot.

text dispersion keyness

Egbert and Biber (2019 524) propose that text dispersion key words 288 can be computed by comparing the number of texts each word is found in in both the study corpus and the reference corpus (instead of comparing word frequencies).

Statistical tests

Three statistical tests are computed:

- Ted Dunning's Log Likelihood test, which measures keyness in terms of the statistical significance and is considered more appropriate than chi-square, especially when contrasting long texts or a whole genre against your reference corpus.
- Log ratio: Andrew Hardie's procedure emphasizing the size of of the keyness as opposed to its statistical significance (related to the %DIFF procedure from Costas Gabrielatos & Anna Marchi but which produces smaller numbers and easier to understand). A value of 2 means the item is 4 times more frequent in the small word list than in the comparison corpus list. A value of 3 means it's 8 times more frequent, and of 4 means it's 16 times more frequent.
- BIC Score. Effectively an alternative to p scores. It uses the log likelihood score and the size of the two corpora in its formula 540. You can leave your p value at 0.1 if you use BIC scores to assess keyness. That will help especially where the comparison corpus is fairly small, as it will tend to bring up more negative key words reflecting the nature of the comparison corpus. Costas Gabrielatos 524 (2018) suggests that BIC scores can be interpreted thus:

below 0	not trustworthy
0-2	only worth a bare mention
2-6	positive evidence
6-10	strong
more than 10	very strong

See UCREL's log likelihood site and SigEff calculator for more on these.

Words get accepted as key if they pass all statistical tests. If you want to ignore the Log ratio test, set its minimum to 0 in the settings 297. If you want to use the BIC score or simply ignore the Log Likelihood test, set the p value 273 to 0.1.

Words which do not occur at all in the comparison corpus are treated as if they occurred 5.0e-324 times (0.0000000 and loads more zeroes before a 5) in such a case. This number is so small as not to affect the calculation materially while not crashing the computer's processor.

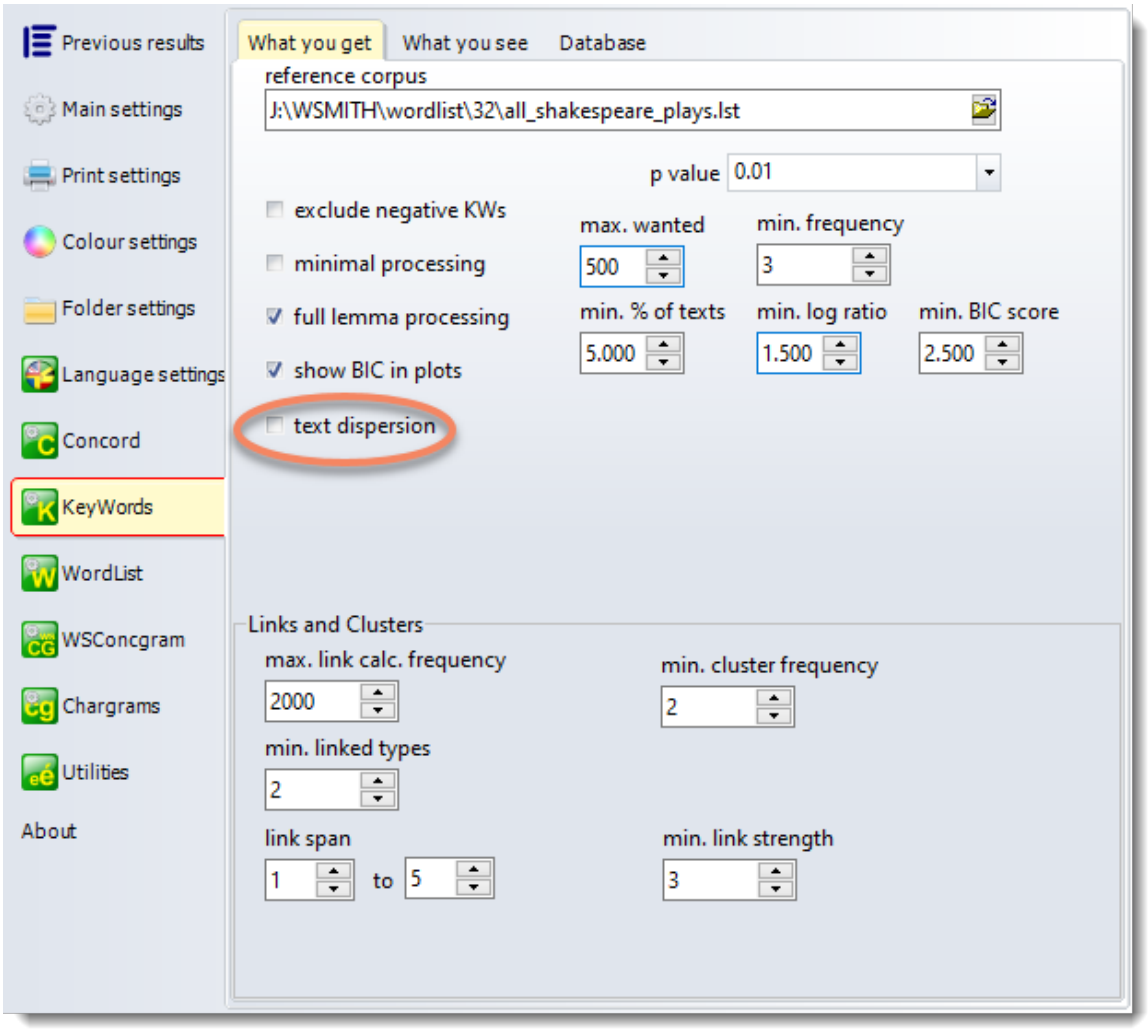
8.9 KeyWords: text dispersion

The point of it

The idea as explained by Egbert and Biber 524 (2019) is to help in the study not of single texts but of sub-corpora of text. They argue that text dispersion key words are more representative of the sub-corpus as a whole.

How to do it

In the Controller settings 297, check the text dispersion box.



Choose your sub-corpus and reference corpus 269 files, remembering that what will count is the number of texts each word is found in.

Example

This was computed using the 24 BNC written academic medicine texts. The reference corpus was the whole BNC XML edition.

J:\WSMITH\KEYWORDS\32K\W medicine_files_XML_text_dispersion.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. Freq.	Rc. %	BIC	Log_L	Log_R
1	E	273	91.67%	22	4	0.05%	203.84	212.16	10.86
2	<adjective>RANDOMISED	164	87.50%	21	181	0.69%	140.73	149.04	6.98
3	<adjective>MYOCARDIAL	212	83.33%	20	238	0.64%	134.37	142.69	7.02
4	<adjective>HISTOLOGICAL	378	83.33%	20	399	0.69%	132.18	140.49	6.91
5	<noun>INFARCTION	237	79.17%	19	269	0.62%	126.90	135.21	7.00
6	MMOL	350	79.17%	19	370	0.67%	124.73	133.04	6.89
7	<adjective>RENAL	419	91.67%	22	478	1.46%	123.54	131.85	5.98
8	<adjective>POSTOPERATIVE	82	79.17%	19	101	0.74%	121.70	130.01	6.74
9	<noun>MORBIDITY	173	91.67%	22	365	1.63%	119.40	127.71	5.81
10	<noun>OUTPATIENT	197	91.67%	22	287	1.75%	116.67	124.98	5.71
11	<adjective>ASYMPTOMATIC	109	75.00%	18	129	0.67%	116.26	124.58	6.81
12	<adjective>ISCHAEMIC	63	70.83%	17	79	0.52%	114.24	122.55	7.09

Percentage columns

In the case of text dispersion key words displays, the *Freq.* column (as before) shows the token frequency of each key word. The *%* column now gives the percentage of *texts* that KW was found in (here a total of 24). The Texts column is as before. The *RC %* column shows the percentage of the number of *texts* each KW was found in.

Results below were derived from the 44 commerce texts:

J:\WSMITH\KEYWORDS\32K\W newsp brdsht nat commerce_files_text_dispersion.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. Freq.	Rc. %	BIC	Log_L	Log_R
1	proper-noun>NIKKEI	39	59.09%	26	97	1.21%	131.65	139.96	5.61
2	<noun>FT-SE	66	52.27%	23	252	1.16%	112.57	120.88	5.49
3	<adjective>PRE-TAX	196	77.27%	34	779	5.11%	108.26	116.57	3.92
4	roper-noun>GOVETT	28	43.18%	19	46	0.74%	99.14	107.46	5.86
5	proper-noun>HANG	22	43.18%	19	52	0.74%	99.14	107.46	5.86
6	<proper-noun>DOW	37	59.09%	26	194	2.91%	93.94	102.25	4.34
7	<proper-noun>SENG	22	43.18%	19	61	0.99%	90.65	98.96	5.45
8	<noun>DIVIDEND	218	77.27%	34	1,454	6.92%	90.62	98.94	3.48
9	roper-noun>BRASIER	25	36.36%	16	28	0.44%	90.11	98.42	6.35
10	<noun>INVESTOR	92	70.45%	31	834	5.46%	89.56	97.88	3.69
11	<noun>TAKEOVER	184	77.27%	34	1,139	8.08%	81.64	89.96	3.26

KWs plot links clusters filenames source text notes

500 entries Row 11 T S Help 327 of 4,049 texts

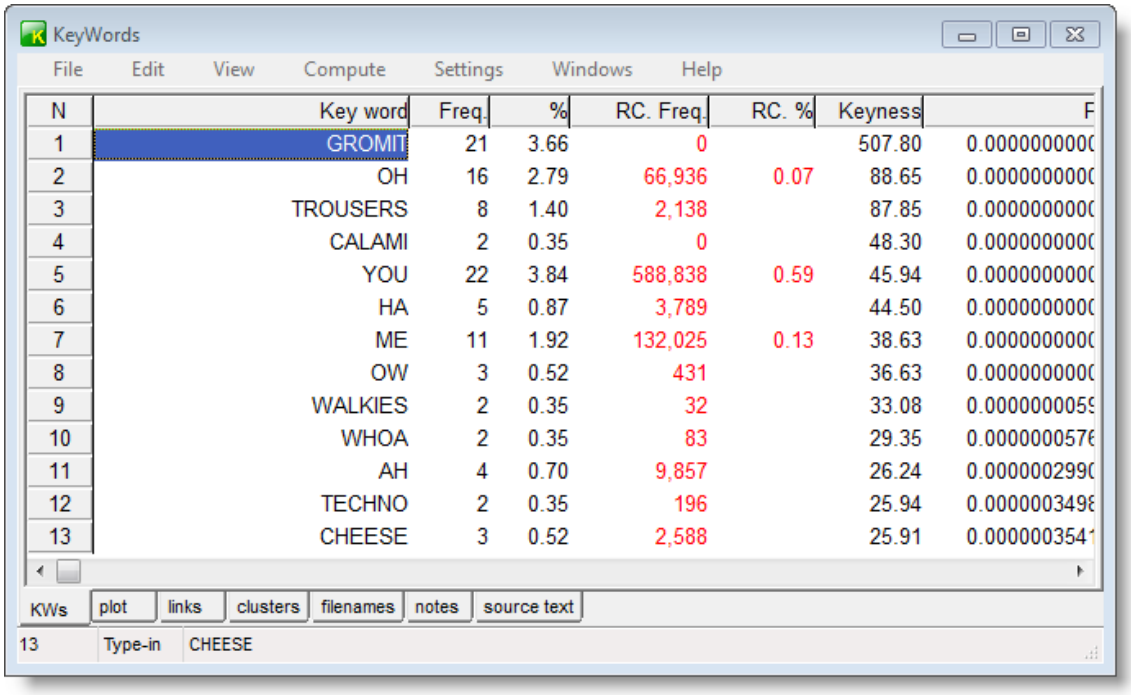
TAKEOVER was found in 34 of 44 commerce texts (77.27%) and the 8.08% represents 327 of the 4,049 reference corpus texts as shown in the status bar.

8.10 KeyWords clusters

What is it?

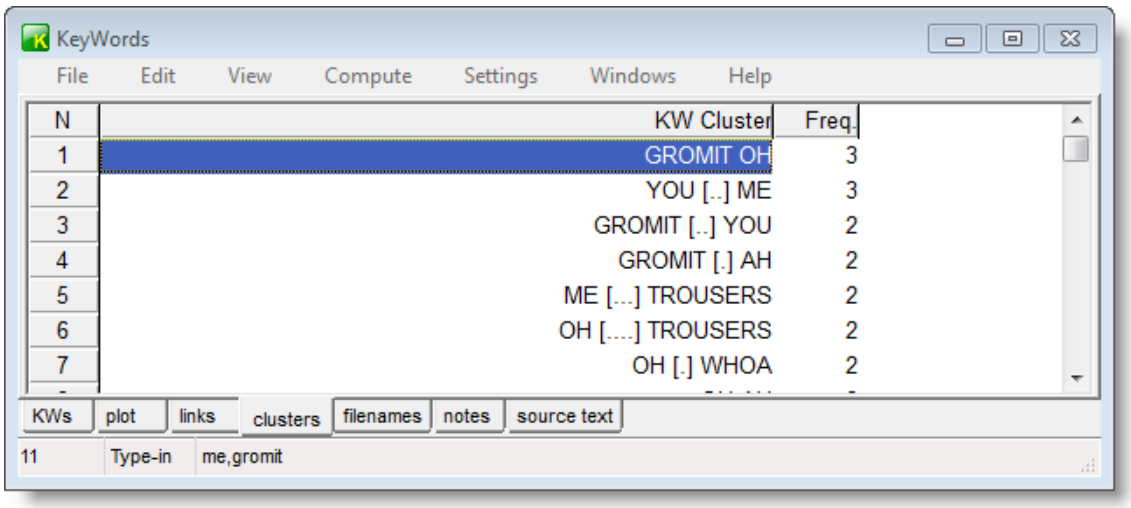
A KeyWords cluster, like a WordList cluster, represents two or more words which are found repeatedly near each other. However, a **KeyWords** cluster only uses key words.

A screenshot will help make things clearer. This is a key words list based on a piece of transcript from a Wallace and Gromit film, using the BNC as the reference corpus.



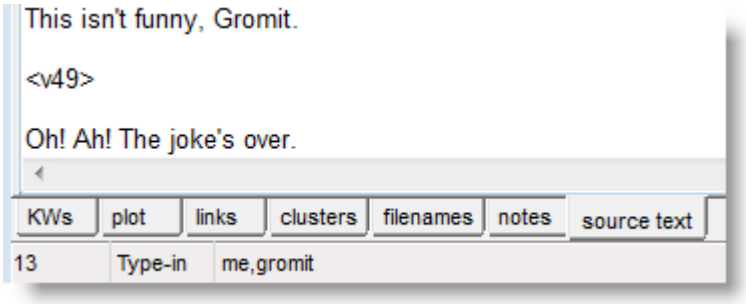
N	Key word	Freq.	%	RC. Freq.	RC. %	Keyness	F
1	GROMIT	21	3.66	0		507.80	0.0000000000
2	OH	16	2.79	66,936	0.07	88.65	0.0000000000
3	TROUSERS	8	1.40	2,138		87.85	0.0000000000
4	CALAMI	2	0.35	0		48.30	0.0000000000
5	YOU	22	3.84	588,838	0.59	45.94	0.0000000000
6	HA	5	0.87	3,789		44.50	0.0000000000
7	ME	11	1.92	132,025	0.13	38.63	0.0000000000
8	OW	3	0.52	431		36.63	0.0000000000
9	WALKIES	2	0.35	32		33.08	0.0000000059
10	WHOA	2	0.35	83		29.35	0.0000000576
11	AH	4	0.70	9,857		26.24	0.0000002990
12	TECHNO	2	0.35	196		25.94	0.0000003498
13	CHEESE	3	0.52	2,588		25.91	0.0000003541

The clusters tab below shows us something like this:



N	KW Cluster	Freq.
1	GROMIT OH	3
2	YOU [..] ME	3
3	GROMIT [..] YOU	2
4	GROMIT [.] AH	2
5	ME [...] TROUSERS	2
6	OH [....] TROUSERS	2
7	OH [.] WHOA	2

The frequency 3 in the **GROMIT OH** line means that there are 3 cases where the key-word **GROMIT** is found within the current collocation span of **OH** in that text. **[.]** means that there is typically one intervening word or [..] two intervening words as in this case shown from the source text.



Requirements

The procedure is text-oriented. You can only get a keywords cluster list if there is exactly one source text. Note that for this procedure sentence boundaries are not blocked, so **Gromit** and **Ah** can be considered to have one word **Oh** intervening.

See also: Plot calculation 296.

8.11 KeyWords: links

The point of it...

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

A plot 296 will show where each key word occurs in the original file. It also shows how many links there are between key-words.

What are links?

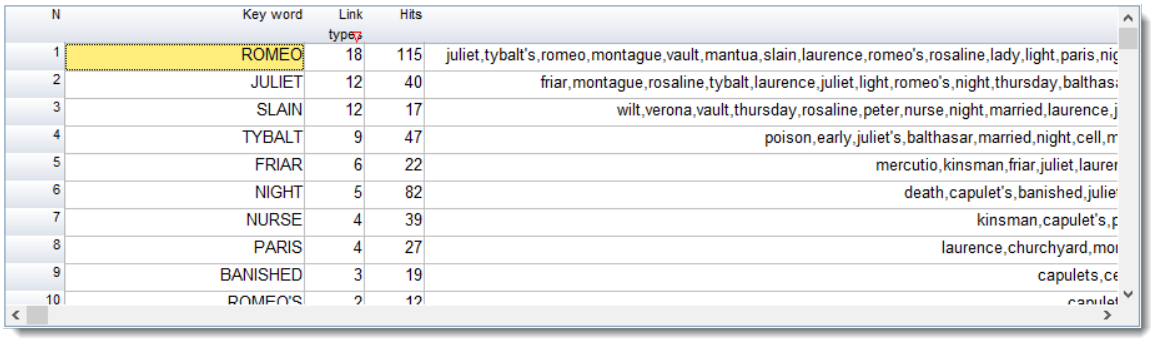
Links are "co-occurrences of key-words within a collocational span". An example is much easier to understand, though:

Suppose the word *elephant* is key in a text about Africa, and that *water* is also a key word in the same text. If *elephant* and *water* occur within a span of 5 words of each other, they are said to be "linked". The number of times they are linked like this in the text will be shown in the Links window.

The link spans (like collocation horizons) go from 1 word away to up to 25 words to left and right. The default 110 is 1 to 5.

What you see

This is a key words list based on Romeo and Juliet, using all the 37 Shakespeare plays as the reference corpus.

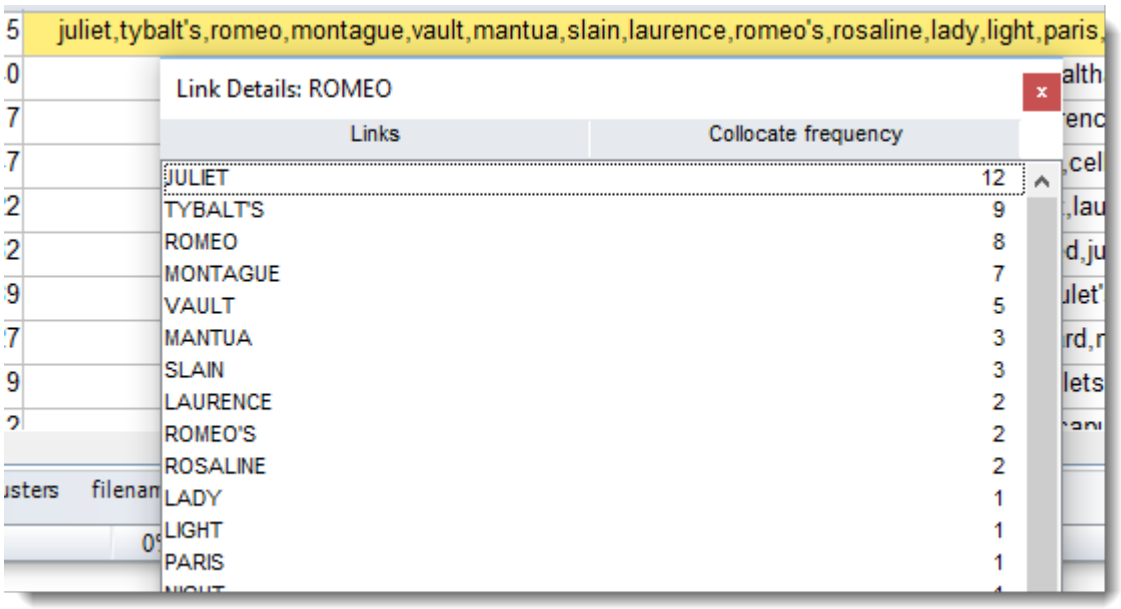


N	Key word	Link types	Hits	
1	ROMEO	18	115	juliet,tybalt's,romeo,montague,vault,mantua,slain,laurence,romeo's,rosaline,lady,light,paris,nig
2	JULIET	12	40	friar,montague,rosaline,tybalt,laurence,juliet,light,romeo's,night,thursday,balthas
3	SLAIN	12	17	wilt,verona,vault,thursday,rosaline,peter,nurse,night,married,laurence,j
4	TYBALT	9	47	poison,early,juliet's,balthasar,married,night,cell,m
5	FRIAR	6	22	mercutio,kinsman,friar,juliet,laure
6	NIGHT	5	82	death,capulet's,banished,julie
7	NURSE	4	39	kinsman,capulet's,p
8	PARIS	4	27	laurence,churchyard,mo
9	BANISHED	3	19	capulets,ce
10	ROMEO'S	2	12	capulet

This Links window shows a number of key words followed by the number of linked types (18 here) the total number of hits of the key word (**ROMEO**) and then the individual linked key words.

Show them all in detail

Double-click any one item in the *Linked KWs* column and you will see its details listed:



ROMEO has 18 linked words; it's linked 18 times with **JULIET**, 9 times with **TYBALT'S**, etc. A right-click menu lets you copy or print these details.

Requirements

The procedure is text-oriented. You can only get a keywords links list if there is exactly one source text.

See also: Plot calculation 296, KeyWords clusters 291, Source Text 134

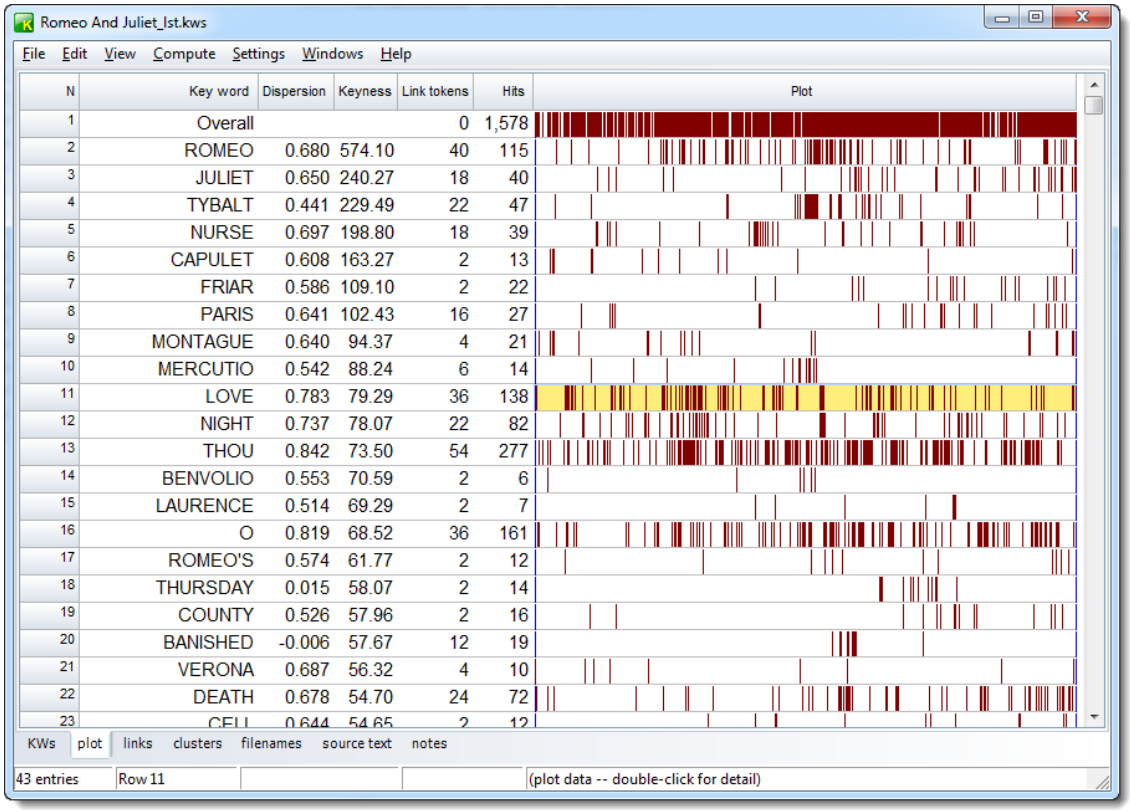
8.12 make a word list from keywords data

With a key word list on your screen, you can press to save your data as a word list (for later comparison, etc. using **WordList** functions).

8.13 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 296 to show which crop up more at the beginning (e.g. in the introduction) and then those from further in the text.

The following screenshot shows KWs of the play *Romeo and Juliet*, revealing where each term occurs. The name **Tybalt**, for example, occurs in a main burst about half way through the text.



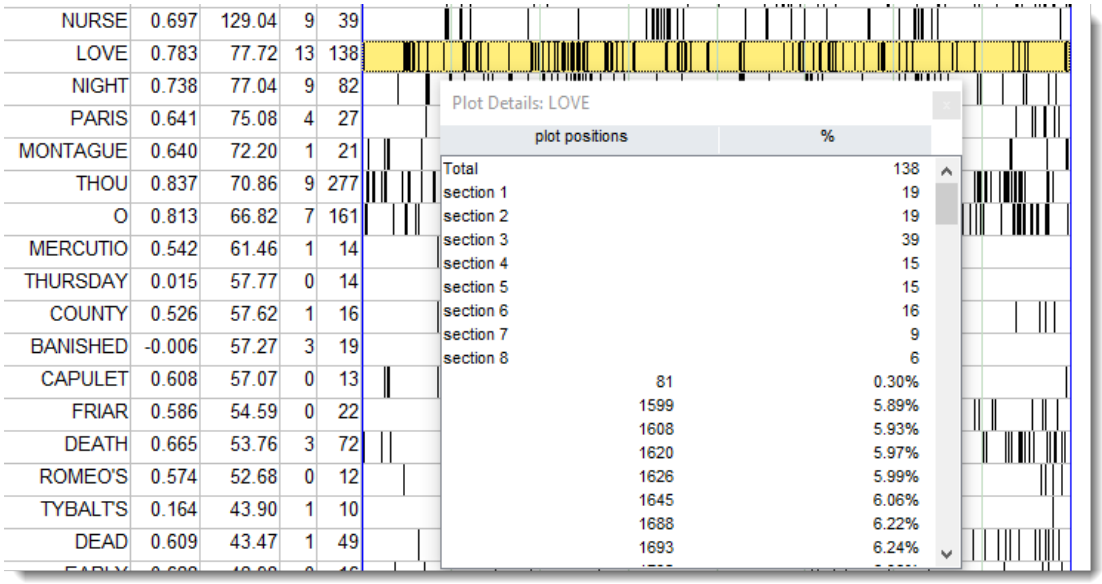
re-sorting

Click the header to re-sort 296 the listing or use the menu **option**. The Key word column sorts alphabetically, the dispersion column sorts on the amount of dispersion (higher numbers mean the occurrences are more spread out); the keyness column is the original plot order, or you can sort on number of links with other KWs or on the number of hits found.

plot data

If the settings 122 permit, KeyWords can compute the context of each hit and show it to you. There is an example from Concord here 230 but the same process operates in KeyWords.

Or you can view KeyWords and Concord plot data as numbers by double-clicking. Here is the view if one double-clicks on the yellow area:



At the top of the list you get total hits and the details of how many hits there are in each ruler segment. Thereafter, the first column gives the word-numbers and the second the percentage of the way through the text. Right-click on this window to copy or print.

links

This shows the total number of links[292] between the key-word and other key-words in the same text, within the current collocation span (default[110] = 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[292] 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[91] option because the plot has graphics, which cannot adequately be represented as text symbols, but you can *Copy* to the clipboard[529] (Ctrl+C) and then paste it into a word processor as a graphic. Alternatively, use the *Output* | *Data as Text File* option, which saves your plot data (each word is followed by the total number of words in the file, then the word number position of each occurrence).

The ruler[548] in the menu (**)** allows you to see the plot divided[116] into equal segments if based on one text, or the text-file divisions if there is more than one.

See also: Key words plot[296], plot dispersion value[554], plot colours and number of divisions[116]

8.13.1 plot calculation

The point of it...

is to see where the key words are distributed within the text. Do they cluster around the middle or near the beginning of the text?

How it's done

This will calculate the inter-relationships between all the key words identified so far, excluding any which you have deleted or zapped 154.

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 554.

Note: this process depends on KeyWords being able to find the source texts 537 which your original word-list was based on.

You may find it useful to export your plot 91 and make other graphs, as explained under Save As 91.

See also: Plot Links 292, Key words plot display 294

8.14 re-sorting: KeyWords

How to do it...

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

the different sorts

A **key words list** offers a choice between sorting by

key-ness (the *key*est words appear at the top)
alphabetical order (from A to Z)
frequency in the smaller list (the most frequent words come first)
frequency in the reference list (the most frequent words come first)

A **key words plot** rotates between sorting by

key-ness (the *key*est words appear at the top)
alphabetical order (from A to Z)
frequency (words which appear oftenest come first)
number of links (the most linked words come first)
first mention of each key word in the text

range (words used in smallest sections of text come first)

A **key key words database** toggles between sorting by

frequency (the most *key key* words appear at the top)

alphabetical order (from A to Z)

An **Associates** 279 **list** toggles between sorting by

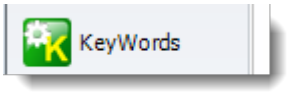
frequency (association between title-word and item)

alphabetical order (from A to Z)

frequency (association between item and title-word)

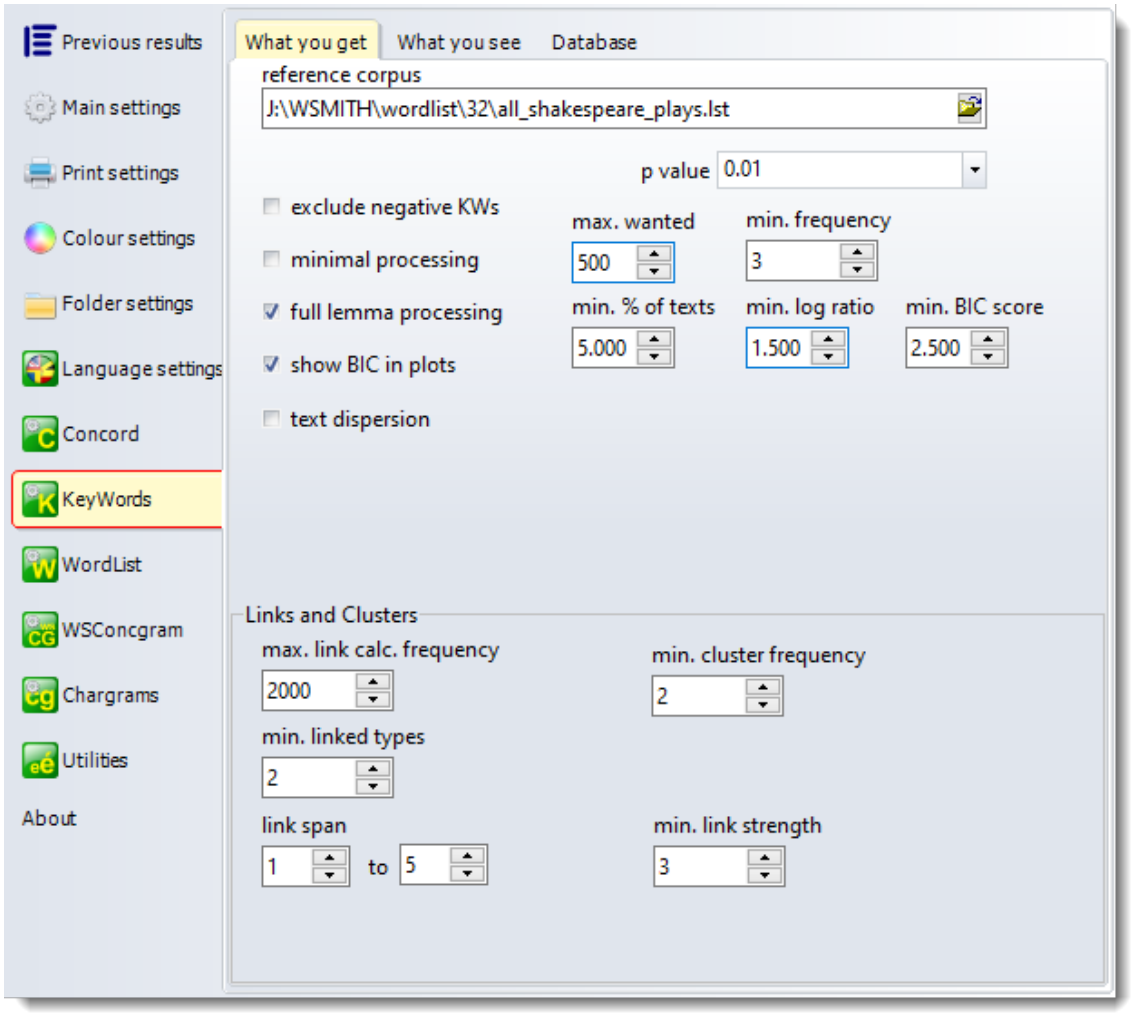
8.15 settings

8.15.1 what you GET



These are found in the main Controller 5 marked *KeyWords.*

This is because some of the choices may affect other Tools. KeyWords and WordList both use similar routines: KeyWords to calculate the key words of a text file, and WordList when comparing word-lists 308.


Previous results
Main settings
Print settings
Colour settings
Folder settings
Language settings
Concord
KeyWords
WordList
WSConcgram
Chargrams
Utilities
About
What you get
What you see
Database
reference corpus
J:\WSMITH\wordlist\32\all_shakespeare_plays.lst
p value 0.01
exclude negative KWs
minimal processing
full lemma processing
show BIC in plots
text dispersion
max. wanted
500
min. frequency
3
min. % of texts
5.000
min. log ratio
1.500
min. BIC score
2.500
Links and Clusters
max. link calc. frequency
2000
min. cluster frequency
2
min. linked types
2
link span
1
to
5
min. link strength
3

WHAT YOU GET

Max. p value

The default level of significance. See p value 273 for more details.

Max. wanted (500), Min. frequency (3), Min. % of texts (5%), Min. log ratio (2.0) and BIC score (2.5)

You may want to restrict the number of key words (KWs) identified so as to find for example the ten most "key" for each text. The program will identify all the key words, sort them by key-ness, and then throw away any excess. It will thus favour positive key words 273 over negative ones.

The minimum frequency is a setting which will help to eliminate any words or clusters which are unusual but infrequent. For example, a proper noun such as the name of a village will usually be extremely infrequent in your reference corpus, and if mentioned only once in the text you're analysing, it is likely not to be "key". The default setting of 3 mentions as a minimum helps reduce spurious hits here. In the case of short texts, less than 600 words long, a minimum of 2 will automatically be used.

The minimum percentage of texts (default = 5%) allows you to ignore words which are not found in many texts. Here the percentage is of the text files in the set you are comparing against a reference corpus. If you're comparing a word-list based on one text, each word in it will occur in 100% of the texts and thus won't get ignored. If you compare a word-list based on 200 texts against your reference corpus, the default of 5% would mean that only words which occur in at least 10 of those texts will be considered for keyness. The KeyWords display 302 shows the number of texts each KW was found in. (If you see ?? that is because the data were computed before that facility came into WordSmith.)

Min. log ratio 288 acts as a threshold (for positive KWs only), so 2.0 omits any KWs which are not at least 4 times as frequent in the text as in your reference corpus.

It is preferable to set a minimum BIC score 288 of at least 2 as explained in the calculation 288 section.

Exclude negative KWs

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

Minimal processing

If this is checked, KeyWords will not compute plots 296, links 292 or KW clusters 291 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, e.g. to make a database.

Full lemma processing

If this is checked, KeyWords will compute the keyness of each lemmatised 322 item. For example if **GO** represents **WENT, GOES** etc. and **GO** alone had a frequency of 10 but the whole set **GO, WENT, GONE** etc. totalled 100,

- if full lemma processing is checked, **GO** would count only 10, and **WENT, GONE** etc will be checked for keyness independently.
- if it is not checked, the frequency of **GO** will be counted as 100 and **WENT, GONE** etc won't get checked for keyness.

show BIC in plots

When a plot is computed, this option will show BIC scores (if checked), otherwise log likelihood scores.

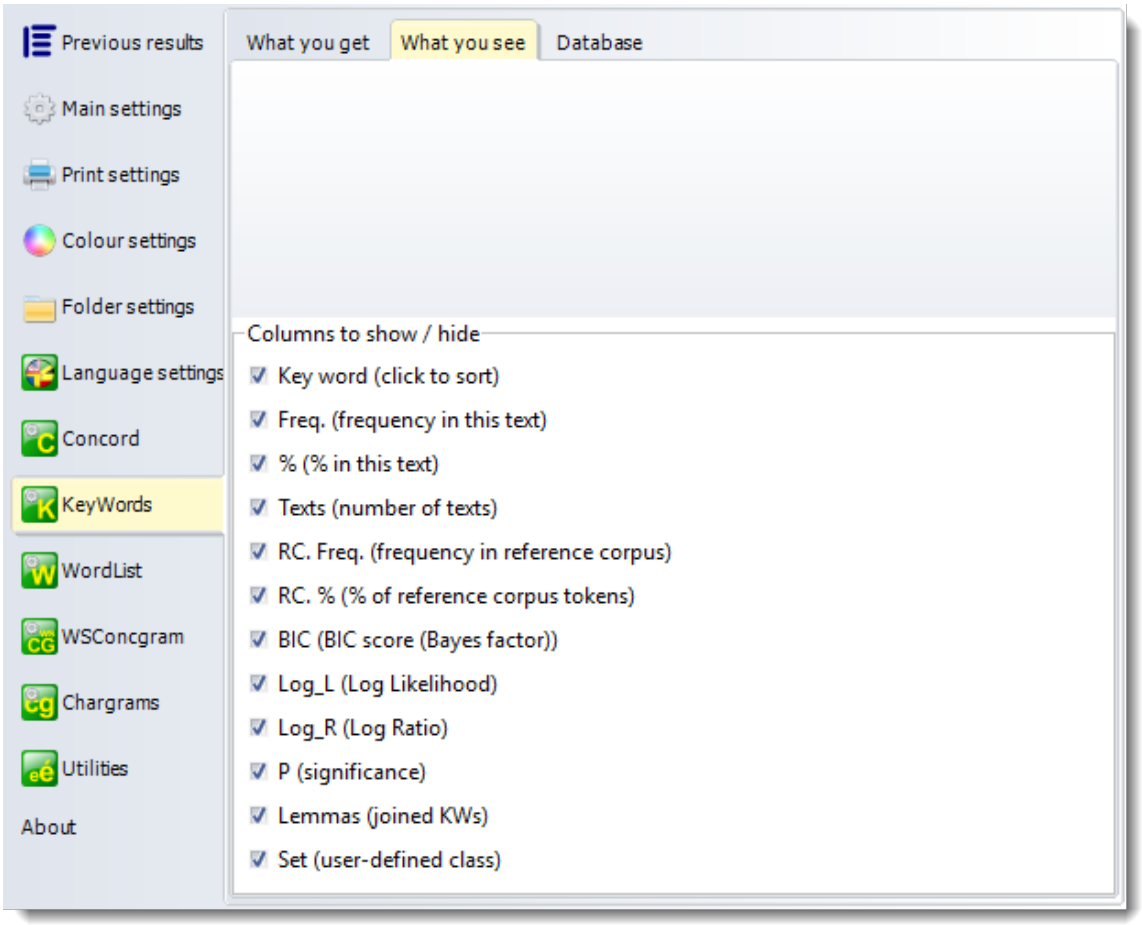
text dispersion

If checked, this computes text dispersion keyness 288.

Links and Clusters

See also: KeyWords Help Contents 266, KeyWords calculation 287.

8.15.2 what you SEE



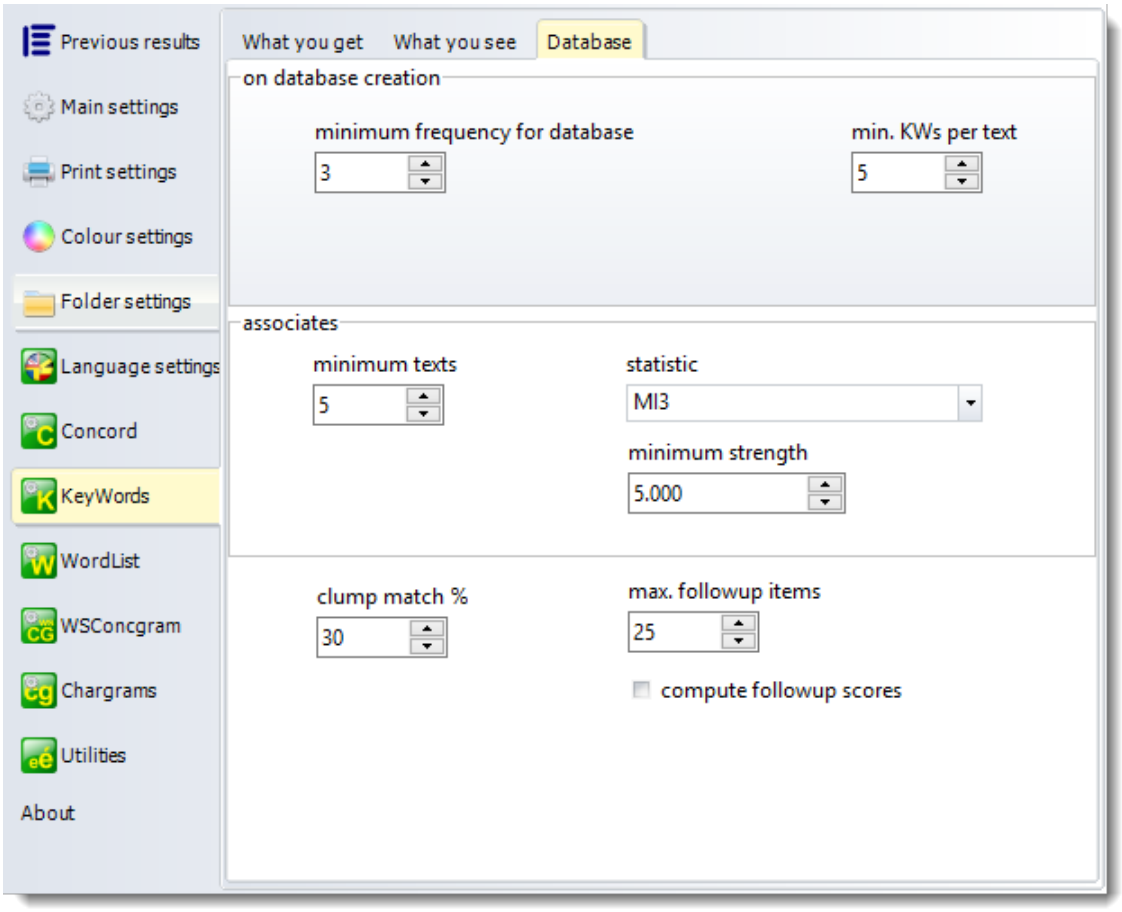
Columns

The *Columns to show/hide* list offers all the standard columns: you may uncheck ones you normally do not wish to see. This will only affect newly computed data: earlier data uses the column visibility, size, colours etc already saved. They can be altered using the Layout 60 menu option at any time.

8.15.3 database settings

The settings are found in the main Controller, *KeyWords* tab, sub-tab *Database*.

Texts here refers to files, assuming 1 file - 1 text.



min. frequency: the default is 10. Any key words which occur in fewer texts won't be included.

min. KWs per text: default = 5. If a set of key words from the same file is smaller it will be ignored,

associates

min. texts: (default = 5). For associates to get computed, they must appear in at least this many texts.

statistic: when computing associates, you can require the linkage between the node term in the database and associates, measured using a particular statistic (such as MI3) to meet or exceed a *minimum strength* (such as 5.0).

clump match %:

max. followup items:

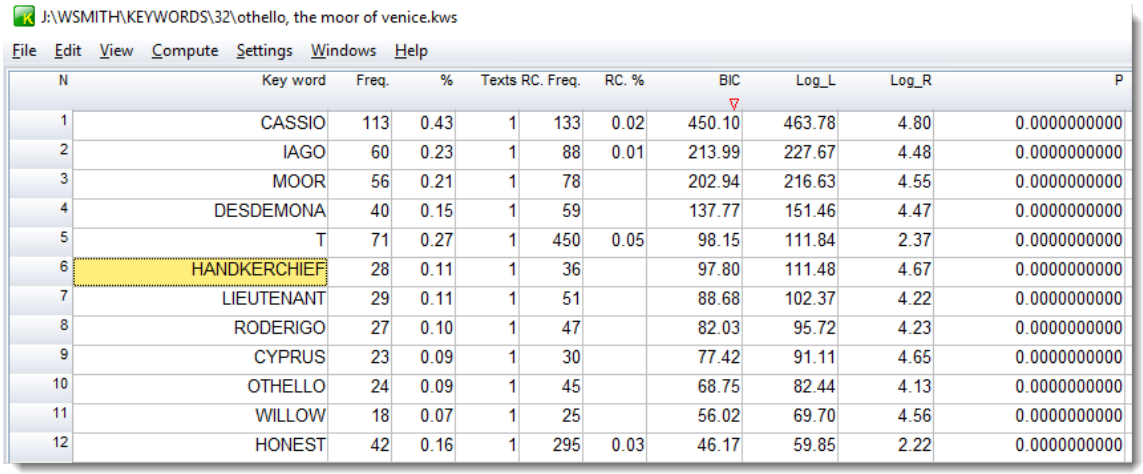
compute followup scores:

See also: database 275, associates 279.

8.16 the key words screen

The display shows

- each key word
- its frequency in the source text(s) which these key words are key in. (*Freq*. column below)
- the % that frequency represents
- the number of texts it was present in
- its frequency in the reference corpus 555 (*RC. Freq*. column)
- the reference corpus frequency as a %
- BIC score 288 indicating of keyness
- Log likelihood statistic 288 of keyness
- Log ratio 288 statistic showing strength of keyness
- p value 273
- lemmas 322 (any which have been joined to each other)
- the user-defined set 43


J:\WSMITH\KEYWORDS\32\othello, the moor of venice.kws

File Edit View Compute Settings Windows Help

N	Key word	Freq.	%	Texts	RC. Freq.	RC. %	BIC	Log_L	Log_R	P
1	CASSIO	113	0.43	1	133	0.02	450.10	463.78	4.80	0.0000000000
2	IAGO	60	0.23	1	88	0.01	213.99	227.67	4.48	0.0000000000
3	MOOR	56	0.21	1	78		202.94	216.63	4.55	0.0000000000
4	DESDEMONA	40	0.15	1	59		137.77	151.46	4.47	0.0000000000
5	T	71	0.27	1	450	0.05	98.15	111.84	2.37	0.0000000000
6	HANDKERCHIEF	28	0.11	1	36		97.80	111.48	4.67	0.0000000000
7	LIEUTENANT	29	0.11	1	51		88.68	102.37	4.22	0.0000000000
8	RODERIGO	27	0.10	1	47		82.03	95.72	4.23	0.0000000000
9	CYPRUS	23	0.09	1	30		77.42	91.11	4.65	0.0000000000
10	OTHELLO	24	0.09	1	45		68.75	82.44	4.13	0.0000000000
11	WILLOW	18	0.07	1	25		56.02	69.70	4.56	0.0000000000
12	HONEST	42	0.16	1	295	0.03	46.17	59.85	2.22	0.0000000000

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 296 according to how outstanding their frequencies of occurrence are. Those near the top are outstandingly frequent.

Negative KWs

At the end of the listing you'll find any which are outstandingly infrequent 273 (negative keywords), in a different colour.

25	BRABANTIO	7	0.03	1	13		24.14	4.14	0.0000008912
26	ALAS	24	0.09	1	229	0.03	23.97	1.78	0.0000009774
27	OUR	51	0.20	1	3,119	0.37	-24.20	-0.90	0.0000008649
28	WE	52	0.20	1	3,394	0.40	-30.83	-1.00	0.0000000252

ALAS is the last ordinary KW: its frequency % of 0.09 is about three times bigger than in the reference corpus (0.03%). **OUR** and **WE** occur at 0.20% in the play Othello but much more (0.37% or

0.40%) in the reference corpus so are considered negatively key: outstandingly infrequent. (The Log Ratio threshold does not apply in the case of negative KWs.)

Limits, Strength, Interpretation

There is no upper limit to the Log_L column of a set of key words. It is not necessarily sensible to assume that the word with the highest log likelihood keyness value must be the most outstanding, since keyness is computed merely statistically; there will be cases where several items are obviously equally key (to the human reader) but the one which is found least often in the reference corpus and most often in the text itself will be at the top of the list.

The log ratio 288 statistic may help you understand the strength of keyness. In this case the word **handkerchief** scores between 4 and 5 which suggests it occurs between 16 and 32 times more often in this play than in the whole set of Shakespeare plays (used as a reference corpus). BIC scores 288 relate Log L to size of the two corpora being compared.

Source text

As its name suggests, choosing the source text tab gets you to a view of the source text 134 (s).

See also : n/a or triangle 384 in the grid

WordList

9 WordList

This program generates word lists based on one or more plain text files. The word lists are automatically generated in both alphabetical and frequency order, and optionally you can generate a word index 330 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 556;
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 77;
5. to get a concordance 69 of one or more of the words in your list.

Within WordList you can compare two lists 308, or carry out consistency analysis (simple 310 or detailed 311) for stylistic comparison purposes.

These word-lists may also be used as input to the KeyWords 266 program, which analyses the words in a given text and compares frequencies with a reference corpus, in order to generate lists of "key-words" and "key-key-words".

Word lists don't have to be of single words, they can be of clusters 332.

See also: WordList display 380

Online step-by-step guide showing how

9.1 index

Explanations

What is WordList and what does it do? 305
Comparing word-lists 308
Comparison Display 309
Consistency Analysis (Simple) 310
Consistency Analysis (Detailed) 311
Definitions 533
Lemmas 322
Limitations 544
Summary Statistics 141
Match List 82

Mutual Information 342
Sort Order 372
Statistics 350
Stop Lists 141
Type/token Ratios 353

Procedures

Auto-Join 326
Batch Processing 30
Calling up a Concordance 69
Choosing Texts 52
Colours 114
Computing a new variable 65

Folders 539
Editing Entries 75
Editing Filenames 152
Keyboard Shortcuts 546
Exiting 91
Fonts 118
Minimum & Maximum Settings 370
Multi-word word lists 332
Mutual Information Score Computing 344
Printing 113
Re-sorting a Word List 372
Saving Results 91
Searching for an Entry by Typing 106
Searching for Entry-types using Menu 341
Single Words or Clusters 332
Text Characteristics 145
Word Index 330
Zapping entries 154

See also: WordSmith Main Index 2, WordList display 380

9.2 clusters

The point of it..

An ordinary word list contains only single words. OK for most words but there are anomalies:

Behind is one word, **in front of** is three.

Happy is one word, **over the moon** three.

New York is two words, **Chicago** only one.

This function lets you provide a list of items you want treated as if they were single items.

How to do it

First you need a plain text list of the phrases you want to treat as single items. Then use the Text Converter 488 to process your corpus seeking out each of these, and marking them. When WordList or Concord processes the text it notices the mark-up and treats items as singles:

35	LOOKING	1	0.93	1	100.00	0.00
36	LOS ANGELES	1	0.93	1	100.00	0.00
37	MAKE	1	0.93	1	100.00	0.00
38	MARKET	1	0.93	1	100.00	0.00
39	MONACO	1	0.93	1	100.00	0.00
40	MOST	2	1.87	1	100.00	0.35
41	NEW YORK	1	0.93	1	100.00	0.00
42	NICER	1	0.93	1	100.00	0.00
43	NOT	1	0.93	1	100.00	0.00
44	OF	6	5.61	1	100.00	0.67

The **<mwu>** cluster tag is a What You Get 367 setting.

See also: cluster lists 332 such as

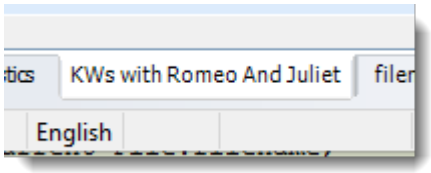
N	Word	Freq.	%	Texts	%	Dispersion	Le...s	Set
1	LOT OF MONEY	37		22	8.15%	0.00		3
2	A LOT OF MONEY	32		19	7.04%	0.00		4
3	A LOT OF	32		19	7.04%	0.00		3
4	OF THE MONEY	29		20	7.41%	0.00		3
5	MONEY OUT OF	23		17	6.30%	0.00		3
6	THE MONEY TO	18		17	6.30%	0.00		3
7	MONEY IN THE	17		15	5.56%	0.00		3
8	AMOUNT OF MONEY	16		14	5.19%	0.00		3
9	VALUE FOR MONEY	15		10	3.70%	0.00		3
10	SUMS OF MONEY	15		13	4.81%	0.00		3

9.3 compare word lists

9.3.1 compute key words

With a word list visible in the WordList tool, you may choose *Compute* | *KeyWords* to get a keywords analysis of the current word list. This will assume you will wish to use the reference corpus 555 defined in the settings 297 for comparison.

You will see the results in one of the tabs at the bottom of the screen.



As in the **KeyWords** tool, this procedure compares all the words in your original word list with those in the reference corpus but does not inform you about words which are only found in the reference corpus.

See also : Compare two wordlists 308, word-list with tags as prefix 374

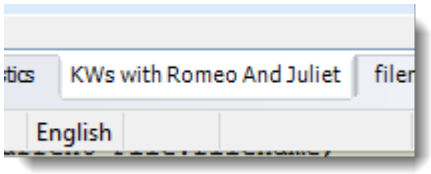
9.3.2 comparing wordlists

The idea is to help stylistic comparisons. Suppose you're studying several versions of a story, or different translations of it. If one version uses *kill* and another has *assassinate*, you can use this function.

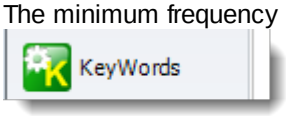
The procedure compares *all the words in both lists* and will report on all those which appear significantly more often in one than the other, including those which appear more than a minimum number of times in one even if they do not appear at all in the other.

How to do it

1. Open a word list.
2. In the menu, choose *File* | *Compare 2 wordlists*.
3. Choose a word list to compare with. You will see the results in one of the tabs at the bottom of the screen.



The minimum frequency (which you can alter in the Controller 5, *KeyWords Settings* tab



) can be set to 1. If it is raised to say 3, the comparison will ignore words which do not appear at least 3 times in at least one of the two lists.

Choose the significance value (all, or a p value 273 from 0.1 to 0.000001 or what you will). The smaller the p value 273, 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 309 format is similar to that used in KeyWords 266. You will also find the Dice coefficient 540 which compares the vocabularies of the two texts, reported in the Notes 28.

See also: Compute Key Words 307, Consistency Analysis 311, Match List 82

9.3.3 comparison display

How to get here? by choosing compare two wordlists 308

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

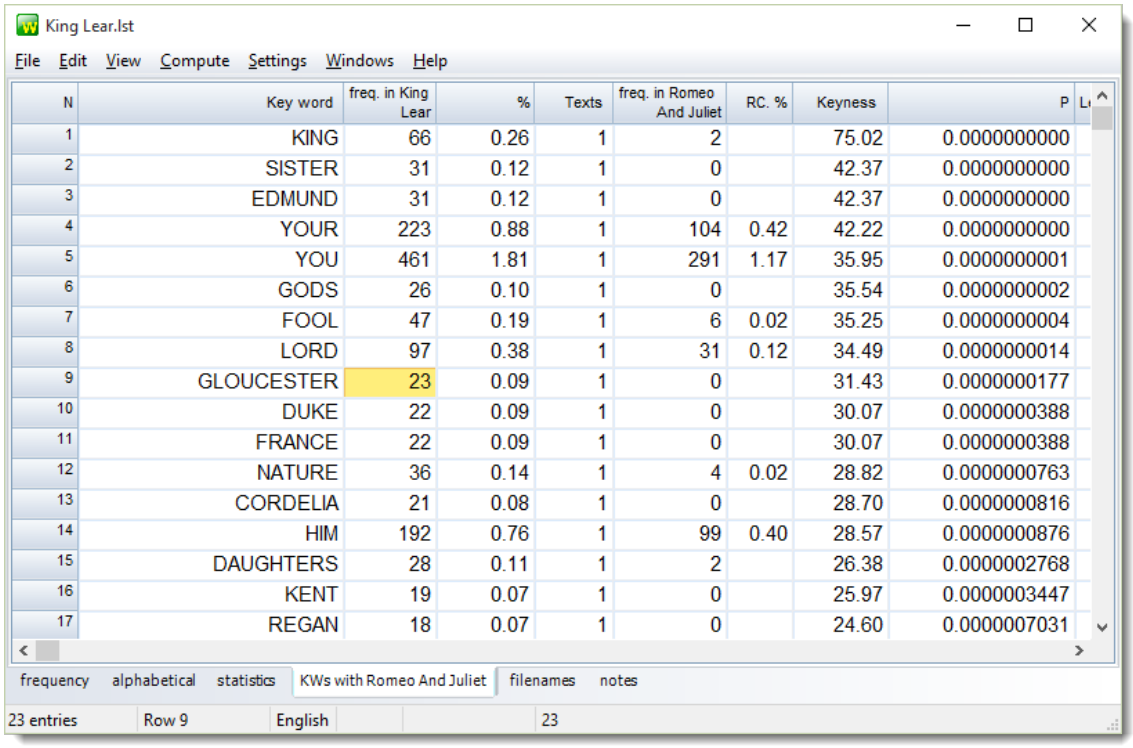
chi-square, log likelihood, or log ratio 287, and p value 273.

The criterion for what counts as "outstanding" is based on the minimum probability value entered before the lists were compared. The smaller this probability value the fewer words in the display.

The words appear sorted according to how outstanding their frequencies of occurrence are. Those near the top are outstandingly frequent in your main word-list. At the end of the listing you'll find those which are outstandingly infrequent in the first text chosen: in other words, key in the second text.

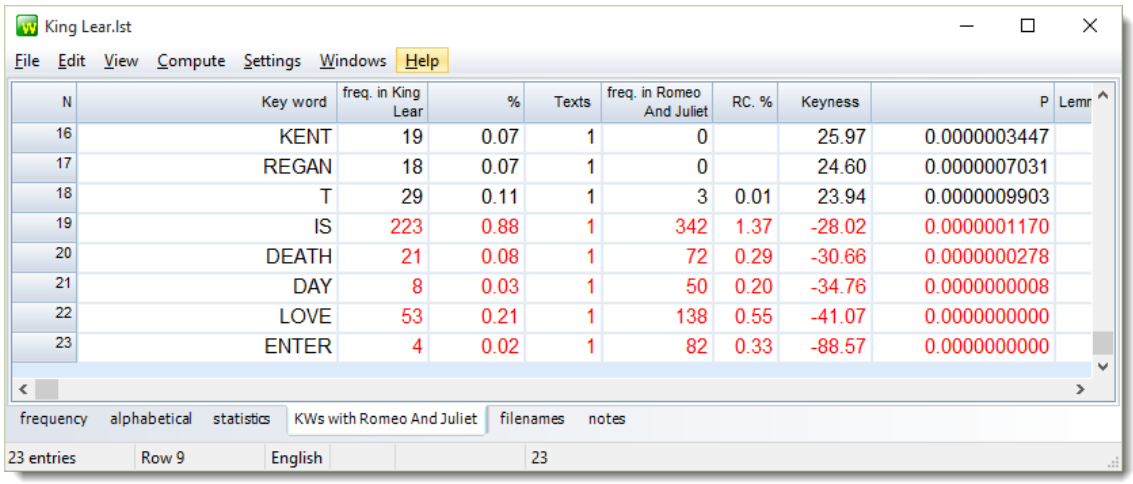
This comparison is similar to the analysis of "key words" in the KeyWords 266 program. The KeyWords analysis is slightly quicker and allows for batch processing.

The word **King** is the most key of all, it scores 75 in the keyness column.



N	Key word	freq. in King Lear	%	Texts	freq. in Romeo And Juliet	RC. %	Keyness	P
1	KING	66	0.26	1	2		75.02	0.0000000000
2	SISTER	31	0.12	1	0		42.37	0.0000000000
3	EDMUND	31	0.12	1	0		42.37	0.0000000000
4	YOUR	223	0.88	1	104	0.42	42.22	0.0000000000
5	YOU	461	1.81	1	291	1.17	35.95	0.0000000001
6	GODS	26	0.10	1	0		35.54	0.0000000002
7	FOOL	47	0.19	1	6	0.02	35.25	0.0000000004
8	LORD	97	0.38	1	31	0.12	34.49	0.0000000014
9	GLOUCESTER	23	0.09	1	0		31.43	0.0000000177
10	DUKE	22	0.09	1	0		30.07	0.0000000388
11	FRANCE	22	0.09	1	0		30.07	0.0000000388
12	NATURE	36	0.14	1	4	0.02	28.82	0.0000000763
13	CORDELIA	21	0.08	1	0		28.70	0.0000000816
14	HIM	192	0.76	1	99	0.40	28.57	0.0000000876
15	DAUGHTERS	28	0.11	1	2		26.38	0.0000002768
16	KENT	19	0.07	1	0		25.97	0.0000003447
17	REGAN	18	0.07	1	0		24.60	0.0000007031

At the bottom we see the words of *King Lear* which are least key in comparison with the play *Romeo and Juliet*.



N	Key word	freq. in King Lear	%	Texts	freq. in Romeo And Juliet	RC. %	Keyness	P	Lemr
16	KENT	19	0.07	1	0		25.97	0.0000003447	
17	REGAN	18	0.07	1	0		24.60	0.0000007031	
18	T	29	0.11	1	3	0.01	23.94	0.0000009903	
19	IS	223	0.88	1	342	1.37	-28.02	0.0000001170	
20	DEATH	21	0.08	1	72	0.29	-30.66	0.0000000278	
21	DAY	8	0.03	1	50	0.20	-34.76	0.0000000008	
22	LOVE	53	0.21	1	138	0.55	-41.07	0.0000000000	
23	ENTER	4	0.02	1	82	0.33	-88.57	0.0000000000	

9.4 merging wordlists

The point of it

You might want to merge 2 word lists (or concordances, mutual information lists etc.) with each other if making each one takes ages or if you are gradually building up a master word list or concordance based on a number of separate genres or text-types.

How to do it

With one word-list (or concordance) opened, choose *File* | *Merge with* and select another.

Be aware that...

Making a merged word list implies that each set of source texts was different. If you chose to merge 2 word lists which both included the same text file, WordSmith will not notice....

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, e.g. in sort order. Presumably if you try to merge one in English with one in Arabic (I've never tried) you should see all the forms but you would get different results merging the Arabic one into the English one (all the Arabic words would be treated as if they were English).

9.5 consistency

9.5.1 consistency analysis (range)

This function (termed "range" by Paul Nation) comes automatically with any word-list.

In any word-list you will see a column headed "Texts". This shows the number of texts each word occurred in (the maximum here being the total number of text-files used for the word-list).

The point of it...

The idea is to find out which words recur consistently in lots of texts of a given genre. For example, the word `consolidate` was found to occur in many of a set of business Annual Reports. It did not occur very often in each of them, but did occur much more consistently in the business reports than in a mixed set of texts.
Naturally, words like `the` are consistent across nearly all texts in English. (While working on a set of word lists to compare with business reports, I found one text without **`the`**. I also discovered that one of my texts was in Italian: but this wasn't the one without **`the`**! The culprit was an election results list, which contained lots of instances of **`Cons., Lab.`** and place names, but no instances of **`the`**.)
To analyse common grammar words like `the`, a consistency list may be very useful. Even so, you're likely to find some common lexical items recur surprisingly consistently.

To eliminate the commonly consistent words and find only those which seem to characterise your genre or sub-genre, you need to find out which are significantly consistent. Save your word list, then use it for comparison[308] with others in WordList, or using KeyWords. This way you can determine which are the significantly consistent words in your genre or sub-genre.

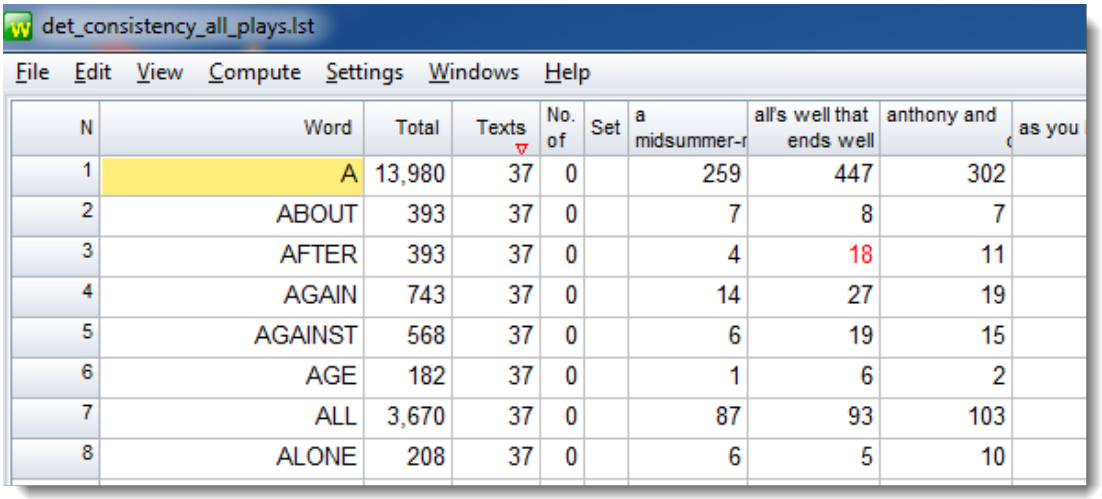
See also: Consistency Analysis (Detailed)[311], Comparing Word-lists[308], Match List[82]

9.5.2 detailed consistency analysis

This function does exactly the same thing as simple consistency[310], but provides much more detail.

The point of it...

The idea is to help stylistic comparisons. Suppose you're studying several versions of a story, or different translations of it. This function enables you to see all the words which are used in the word lists which you have called up.



N	Word	Total	Texts	No. of	Set	a midsummer-r	all's well that ends well	anthony and	as you
1	A	13,980	37	0		259	447	302	
2	ABOUT	393	37	0		7	8	7	
3	AFTER	393	37	0		4	18	11	
4	AGAIN	743	37	0		14	27	19	
5	AGAINST	568	37	0		6	19	15	
6	AGE	182	37	0		1	6	2	
7	ALL	3,670	37	0		87	93	103	
8	ALONE	208	37	0		6	5	10	

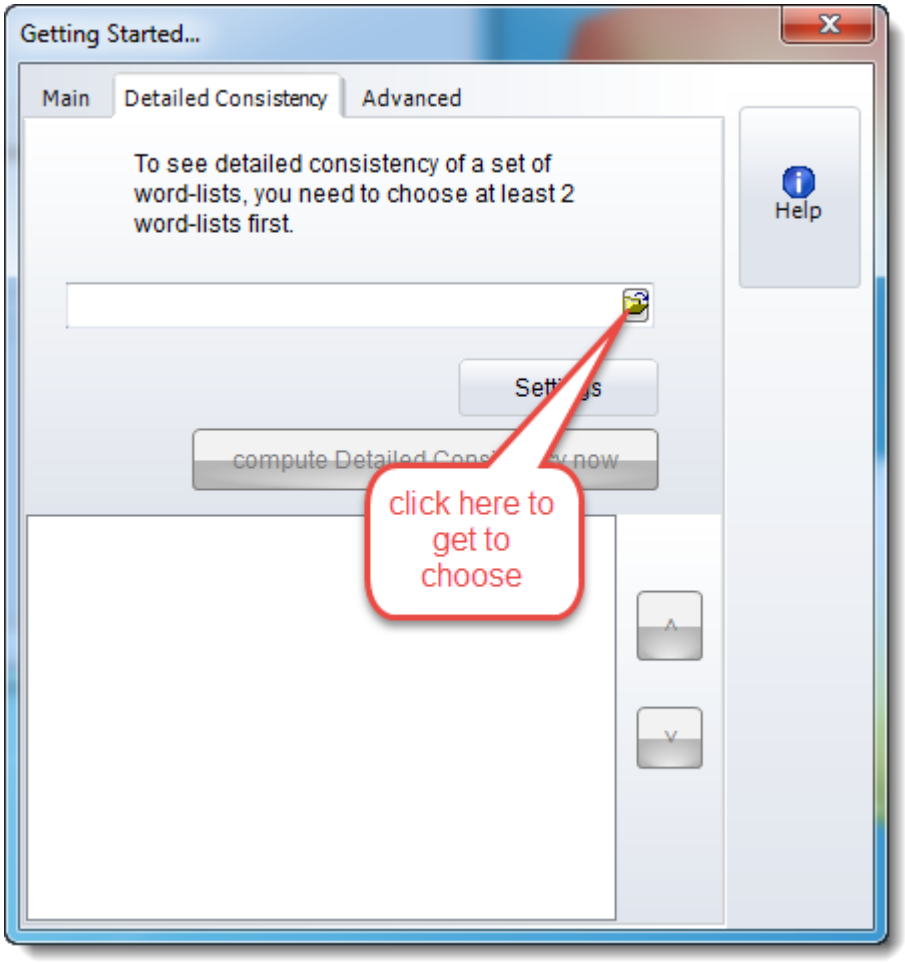
The Total column shows how many instances of each word occurred overall, Texts shows how many text-files it came in. Then there are two columns (No. of Lemmas, and Set which behaves as in a word-list) and then a column for each text. In this case, the word **`after`** occurred in all 37

texts, it occurred 393 times in all, and it was most frequent in **all's well that ends well** at 18 occurrences. Statistics and filenames can be seen for the set of 37 Shakespeare plays used here by clicking on the tabs at the bottom. Notes 28 can be edited and saved along with the detailed consistency list.

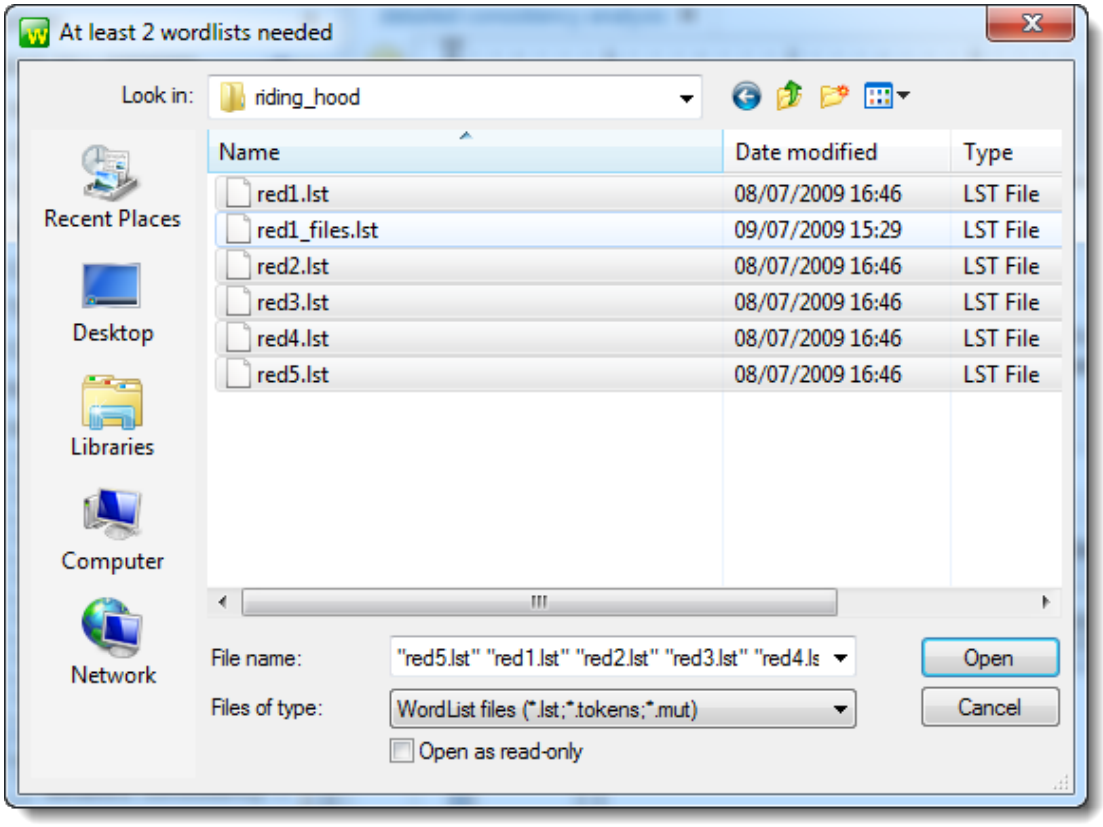
There is no limit except the limit of available memory as to how many text files you can process in this procedure. You can set a minimum number of texts and a minimum overall frequency in the WordList settings in the Controller 367.

How to do it...

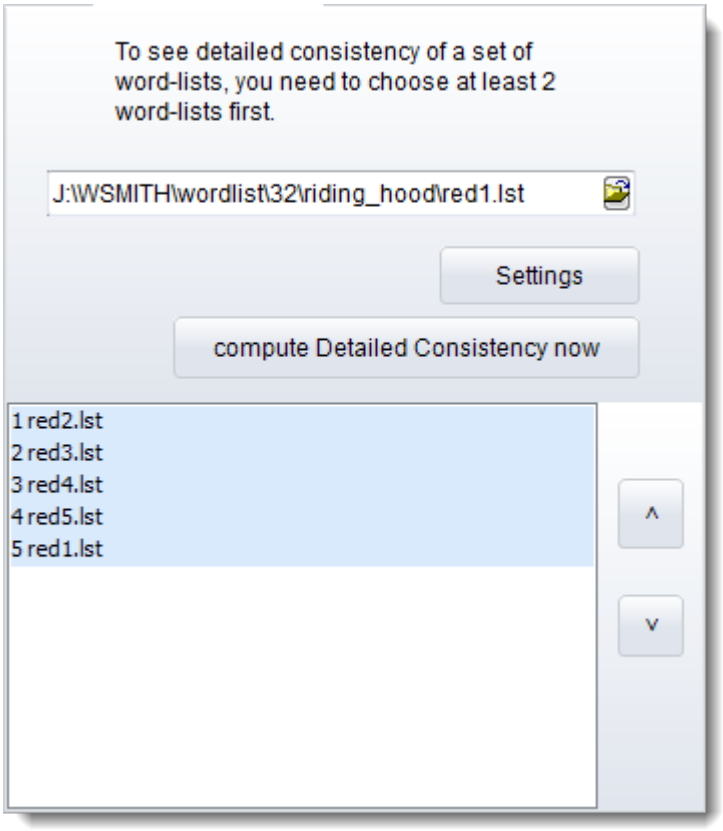
In the window you see when you press **New...(** **)** you will be offered a tab showing detailed consistency.



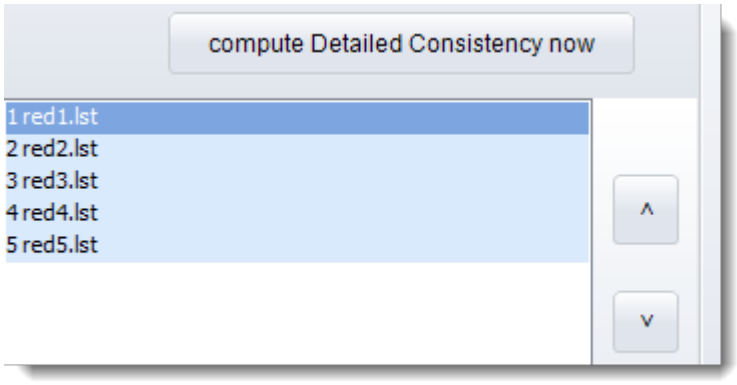
To choose more than 1, use Control or Shift as you click. Below I have chosen five out of 6 available. (These are versions of Red Riding Hood.)



Initially they may come in the wrong order:



so adjust with the two buttons at the right.



and now press **compute Detailed Consistency now**.

Settings

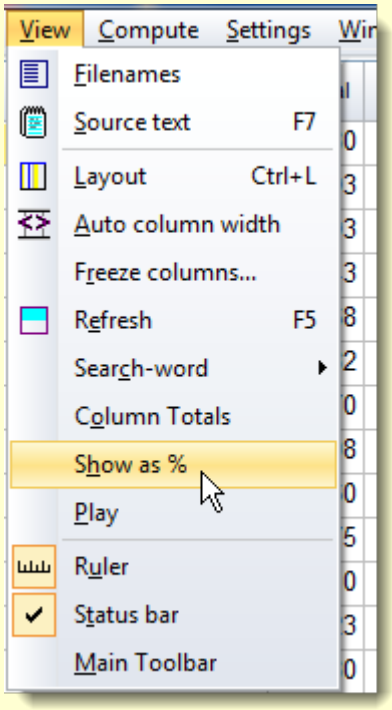
You can require a minimum number of texts and minimum frequency in the main Controller 367 if you click this.

Sorting

Each column can be sorted by clicking on its header column (**Word, Freq.** etc.). When working on Shakespeare plays, to get the words which occurred in all 37 to the top, I clicked **Texts**.

Row percentages

If you choose to *Show as %*,



you can change this view:

Detailed Consistency List (J:\WSMITH\wordlist\32\consistency_(Hamlet Lear Romeo).lst)

File Edit View Compute Settings Windows Help

N	Word	Total	Texts	No. of	Set	Hamlet, prince of	Romeo and juliet	King lear
1	THE	2,632	3	0		1,099	658	875
2	AND	2,239	3	0		901	664	674
3	TO	1,843	3	0		747	551	545
4	I	1,783	3	0		562	579	642
5	OF	1,505	3	0		660	384	461

to this:

Detailed Consistency List (J:\WSMITH\wordlist\32\consistency_(Hamlet Lear Romeo).lst)

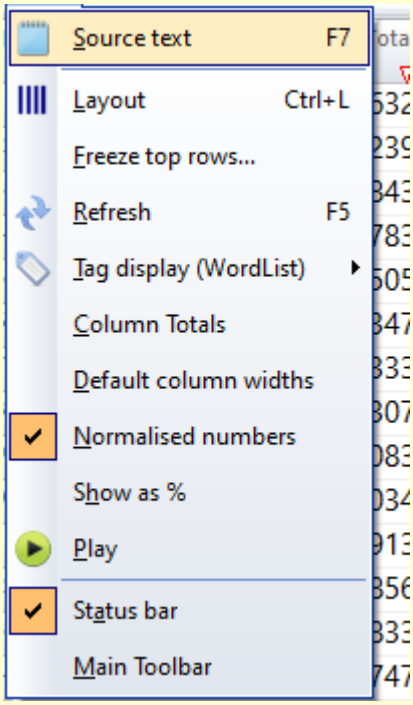
File Edit View Compute Settings Windows Help

N	Word	Total	Texts	No. of	Set	Hamlet, prince of	Romeo and juliet	King lear
1	THE	2,632	3	0		41.8%	25.0%	33.2%
2	AND	2,239	3	0		40.2%	29.7%	30.1%
3	TO	1,843	3	0		40.5%	29.9%	29.6%
4	I	1,783	3	0		31.5%	32.5%	36.0%
5	OF	1,505	3	0		43.9%	25.5%	30.6%

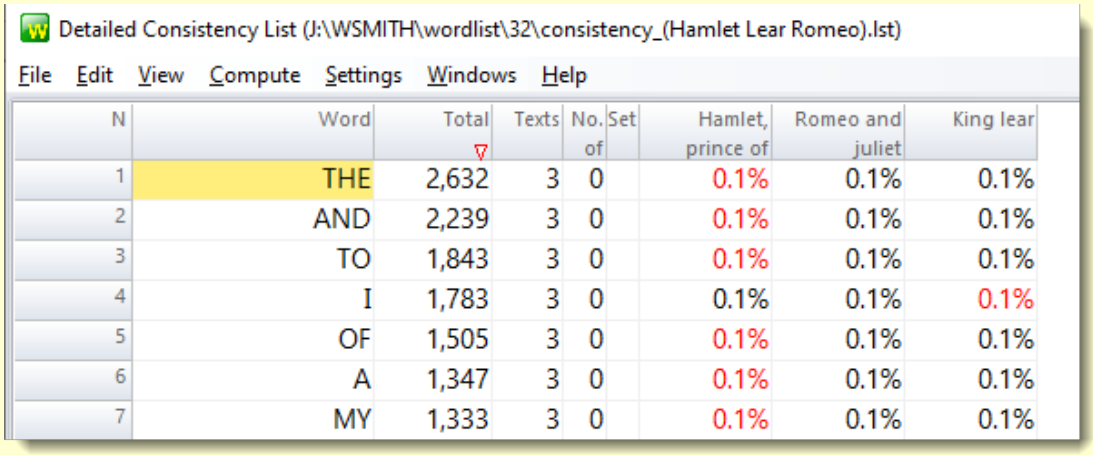
with the percentages adding to 100 across each row.

Normalised counts

If you choose *View* | *Normalised numbers*,



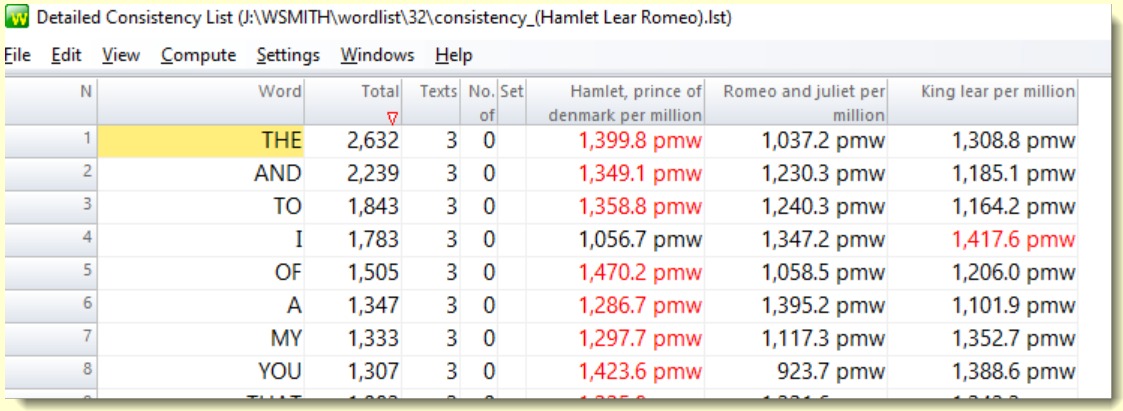
you will get a view like this:


Detailed Consistency List (J:\WSMITH\wordlist\32\consistency_(Hamlet Lear Romeo).lst)

File Edit View Compute Settings Windows Help

N	Word	Total	Texts	No. of Set	Hamlet, prince of	Romeo and juliet	King lear
1	THE	2,632	3	0	0.1%	0.1%	0.1%
2	AND	2,239	3	0	0.1%	0.1%	0.1%
3	TO	1,843	3	0	0.1%	0.1%	0.1%
4	I	1,783	3	0	0.1%	0.1%	0.1%
5	OF	1,505	3	0	0.1%	0.1%	0.1%
6	A	1,347	3	0	0.1%	0.1%	0.1%
7	MY	1,333	3	0	0.1%	0.1%	0.1%

where the numbers take into account the number of tokens in each text file. However those numbers are tiny and all look the same, so you will probably want to adjust the numbers to per thousand or per million words[131].


Detailed Consistency List (J:\WSMITH\wordlist\32\consistency_(Hamlet Lear Romeo).lst)

File Edit View Compute Settings Windows Help

N	Word	Total	Texts	No. of Set	Hamlet, prince of denmark per million	Romeo and juliet per million	King lear per million
1	THE	2,632	3	0	1,399.8 pmw	1,037.2 pmw	1,308.8 pmw
2	AND	2,239	3	0	1,349.1 pmw	1,230.3 pmw	1,185.1 pmw
3	TO	1,843	3	0	1,358.8 pmw	1,240.3 pmw	1,164.2 pmw
4	I	1,783	3	0	1,056.7 pmw	1,347.2 pmw	1,417.6 pmw
5	OF	1,505	3	0	1,470.2 pmw	1,058.5 pmw	1,206.0 pmw
6	A	1,347	3	0	1,286.7 pmw	1,395.2 pmw	1,101.9 pmw
7	MY	1,333	3	0	1,297.7 pmw	1,117.3 pmw	1,352.7 pmw
8	YOU	1,307	3	0	1,423.6 pmw	923.7 pmw	1,388.6 pmw

See also: Detailed Consistency Relations 319, Consistency Analysis (range) 310, Comparison Display 309, Comparing Word-lists 308, Match List 82, Column Totals 67

9.5.2.1 re-sorting: consistency lists

The frequency-ordered consistency display can be re-sorted by
alphabetical order (Word)
total frequencies overall (Total, the default)
by the *frequencies* in any given file (you see the file names).
Click on Word, Total or a filename to choose.
The sort can be either ascending or descending, the default being descending.

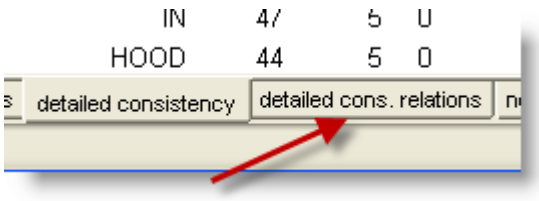
See also: Sorting word-lists 372

9.5.3 detailed consistency relations

With a detailed consistency list 311 such as this, of five versions of the fairy story *Little Red Riding Hood*,

N	Word	Total	Texts	as	Set	red1	red2	red3	red4	red5
4	HER	100	5	0		8	22	26	6	38
5	SHE	96	5	0		5	14	20	5	52
6	A	83	5	0		12	12	15	9	35
7	YOU	70	5	0		14	15	22	1	18
8	RED	66	5	0		1	13	20	1	31
9	OF	66	5	0		14	8	8	4	32
10	WAS	56	5	0		3	9	9	4	31
11	GRANDMOTHER	53	4	0		0	12	17	5	19
12	LITTLE	51	5	0		7	16	17	7	4
13	IN	47	5	0		5	7	10	3	22
14	HOOD	44	5	0		1	13	1	1	28

it looks as if the most long-winded story is probably version 5 (**`red5.lst`**). If you click the *detailed cons. relation* tab



you can see the relevant statistics more usefully:

N	File 1	Count	File 2	Count	Joint	Relation	Set
1	red1.lst	169	red2.lst	234	83	0.412	
2	red1.lst	169	red3.lst	333	89	0.355	
3	red1.lst	169	red4.lst	98	42	0.315	
4	red1.lst	169	red5.lst	462	89	0.282	
5	red2.lst	234	red3.lst	333	138	0.487	
6	red2.lst	234	red4.lst	98	57	0.343	
7	red2.lst	234	red5.lst	462	136	0.391	
8	red3.lst	333	red4.lst	98	61	0.283	
9	red3.lst	333	red5.lst	462	162	0.408	
10	red4.lst	98	red5.lst	462	61	0.218	

where it can be seen that **red5** has a type-count of 462 words, more than any other, and that the relation between **red2** and **red3** is the closest with a relation statistic of 0.487.

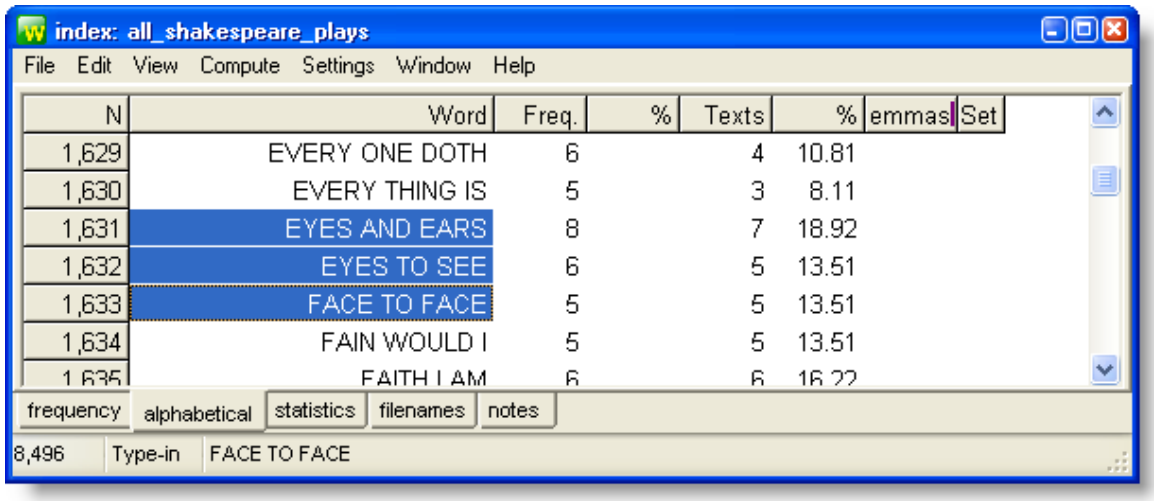
This relation is the Dice coefficient 540, based on the joint frequency and the type-counts of the two texts. Type count is the number of different word types 533 in each text. Joint frequency: there are 138 matches in the vocabulary of these two versions, which means that 138 distinct word types matched up in the two word lists. (If for example **book** appeared 20 times in one list and 3 times in the other, that would count as 1 match.)

A Dice coefficient ranges between 0 and 1. The 0.487 can be thought of like a percentage, i.e. there's about a 49% overlap between the vocabularies of the two versions of the same story.

See also : Detailed Consistency 311.

9.6 find filenames

If you have an index-based word list on screen you can see how many text files each word was found in. For example, in this index based on Shakespeare plays, **EYES AND EARS** occurs in 7 of the 37 plays.



What if you want to know *which* of those plays?

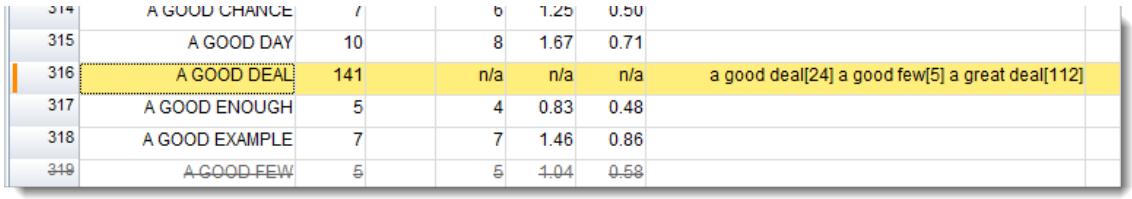
Select the word(s) or cluster(s) you're interested in and choose *File* | *Find Files* in the menu and you will get something like this:

See also : source texts 134, selecting multiple entries 107, making a WordList index 330, n/a or triangle 384 in display

9.7 Lemmas (joining words)

In a word list, a key word list or a list of collocates you may want to store several entries together: e.g. **want; wants; wanting; wanted.** Bringing them together means you're treating them as members of the same "lemma" or set -- rather like a headword in a dictionary.

A lemmatised head entry has a coloured 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.



314	A GOOD CHANCE	7		6	1.25	0.50	
315	A GOOD DAY	10		8	1.67	0.71	
316	A GOOD DEAL	141		n/a	n/a	n/a	a good deal[24] a good few[5] a great deal[112]
317	A GOOD ENOUGH	5		4	0.83	0.48	
318	A GOOD EXAMPLE	7		7	1.46	0.86	
319	~~A GOOD FEW~~	~~5~~		~~5~~	~~1.04~~	~~0.58~~	

Here we see a word list based on 3-word clusters 332 where originally **a good deal** had a frequency of 24, but has been joined to **a great deal** and **a good few** and thereby risen to 141.

You'll see **n/a** 384 in some columns because if you join after a word list was computed some information is no longer available. Fix it with a lemma consistency check 324.

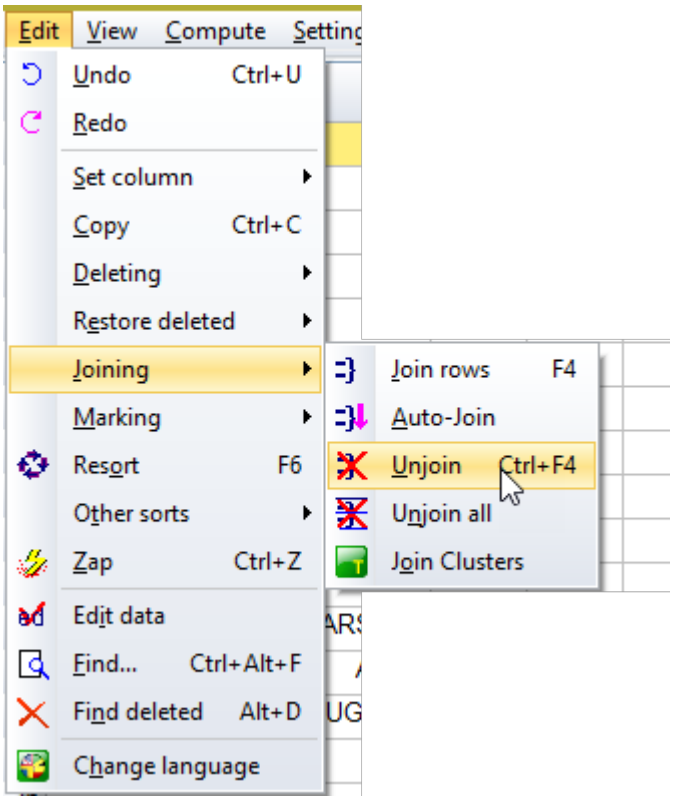
See also: joining words 322, manual joining 325, auto-joining 326, choosing lemma file 327

9.7.1 how do we join words?

Joining can be done automatically 326 or manually 325. As explained here 324, dispersion 354 is not shown for words which have been lemmatised after the word list was computed.

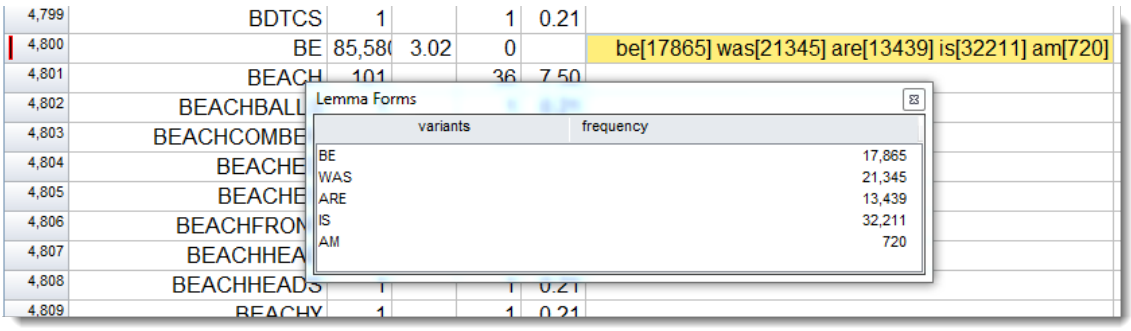
How to do it

The menu offers choices for joining and un-joining.



View all the various lemma forms

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

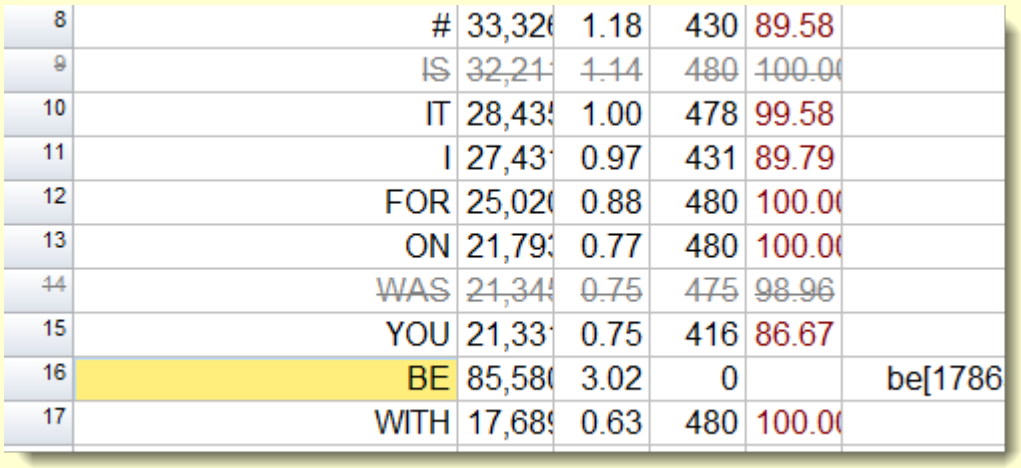


and a window of Lemma Forms will open up, showing the various components.

Advanced

Get rid of the deleted words

If you don't want to see the deleted words

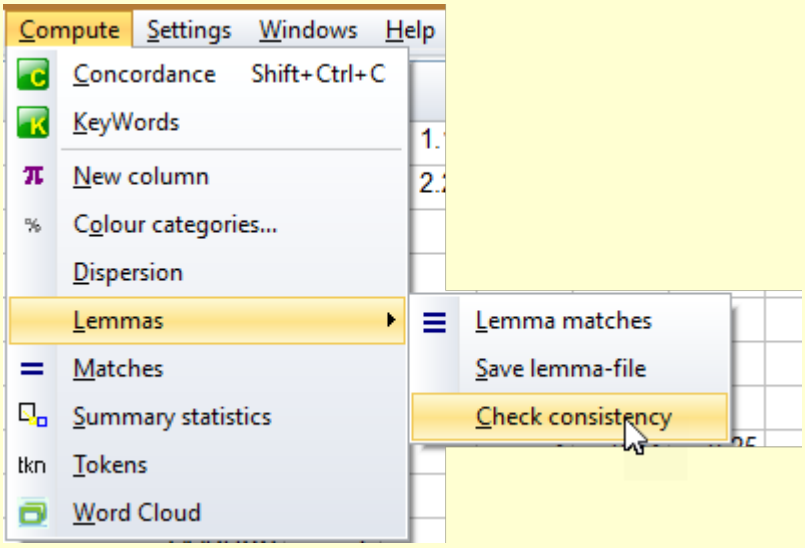


8	#	33,32	1.18	430	89.58	
9	~~IS~~	~~32,21~~	~~1.14~~	~~480~~	~~100.0~~	
10	IT	28,43	1.00	478	99.58	
11	I	27,43	0.97	431	89.79	
12	FOR	25,02	0.88	480	100.0	
13	ON	21,79	0.77	480	100.0	
14	~~WAS~~	~~21,34~~	~~0.75~~	~~475~~	~~98.96~~	
15	YOU	21,33	0.75	416	86.67	
16	BE	85,58	3.02	0		be[1786
17	WITH	17,68	0.63	480	100.0	

choose Ctrl-Z to zap 154 them.

How many text-files?

When you join or un-join words the number of text files for a given entry becomes unknown and you'll see a blank or 0 in the *Texts* column of a word list as in the case of **BE** above. (This is because although WordList does know how many texts each of the lemma variants occurs in, it doesn't know how many overlap.) To determine how many belong to the whole set, choose *Check consistency*.



To do it, WordList needs to be supplied with the name of a suitable index file 330 based on the same texts.

See also: Auto-Joining methods 326, Using a text file to lemmatise 327, selecting multiple entries 107, Concord lemmatisation 217, Relative entropy 356

9.7.2 manual joining

Manual joining

You can simply do this by dragging one entry to another. Suppose your word list has

WANT
WANTED
WANTING

you can simply grab **wanting** or **wanted** with your mouse and place it on **want**.

(See choosing lemma file 327 if you want to join these to a word which isn't in the list)

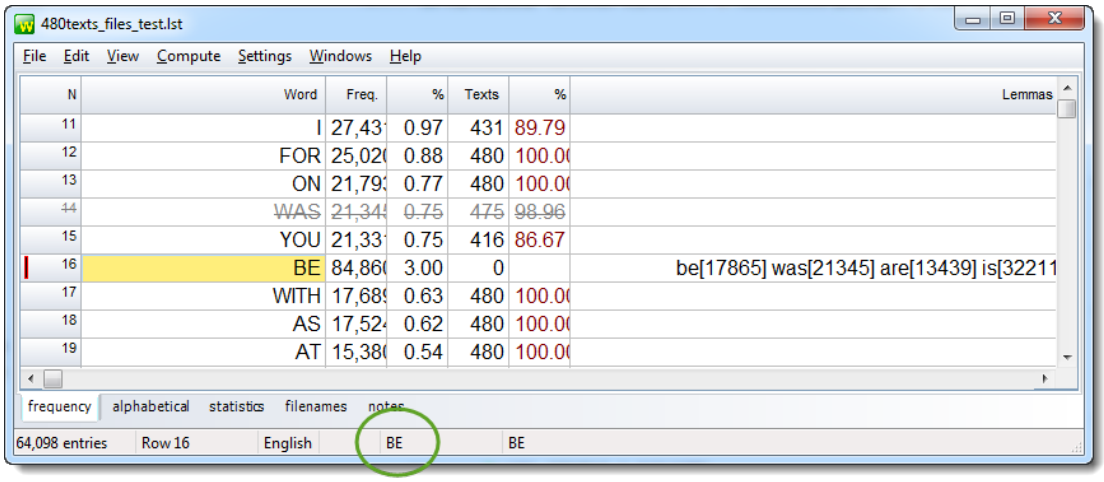
Can't see the word to join to?

If you cannot see all the items you want to join in one screen, you can do the same thing using by marking 108.

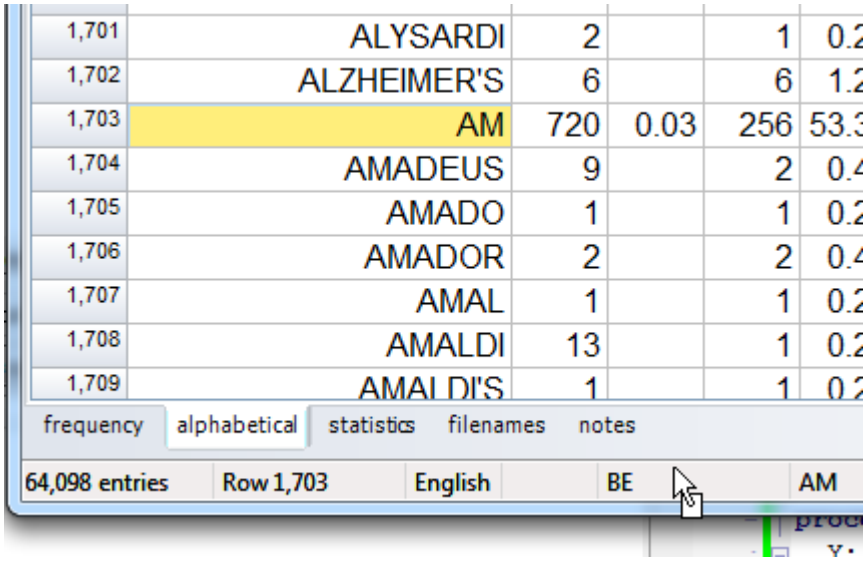
1. Use Alt+F5 to mark an entry for joining to another. The first one you mark will be the "head". For the moment, while you're still deciding which other entries belong with it, the edge of that row will be marked green. Any entries which you then decide to link with the head (by again pressing Alt+F5) will show they're marked too, in white. (If you change your mind you can press Shift+Alt+F5 and the marking will disappear.)

2. Use F4 to join all the entries which you've marked. The program will then put the joint frequencies of all the words you've marked with the frequency of the one you marked 108 first (the head).

Alternatively, 1. select the head word, this makes it visible in the status bar.



2. Find the word you want to join and drag it to the status bar.



To Un-join

If you select an item which has lemmas visible at the right and press Ctrl+F4, this will unjoin the entries of that one lemma. To unjoin all lemmatised forms in the entire list, in the menu choose *Edit | Join | Unjoin All*.

9.7.3 auto-joining lemmas

There are two methods, a) based on a list, and b) based on a template.

a) File-based joining

You can join up lemmas using a text file 327 which automates the matching & joining process. The actual processing of the list takes place when you choose the menu option *Match Lemmas* ( ) in WordList, Concord or KeyWords. Every entry in your lemma list will be checked to see whether it matches one of the entries in your word list. In the example, if, say, ***am, was***, and ***were*** are found, they will be stored as lemmas of ***be***. If ***go*** and ***went*** are found, then ***went*** will be joined to ***go***.

b) Auto-joining based on a template

Or you can auto-join any of the entries in your current word list which meet your criteria: the menu option *Auto-Join* can be used to specify a string such as **S** or **S;ED;ING** and will then go through the whole word list, lemmatising all entries where one word only differs from the next by having **S** or **ED** or **ING** on the end of it. (Use **;** to separate multiple suffixes.)

Prefix / Suffix / Infix

By default all strings typed in are assumed to be suffixes; to join prefixes put an asterisk (*****) at the right end of the prefix. If you want to search for infixes (eg. **bloody** in **absobloodylutely** [languages like Swahili use infixes a lot]) put an asterisk at each end.

Examples

S;ED;ING will join **books** to **book**, **booked** to **book** and **booking** to **book**

***S;*ED;*ING** will join **books** to **book, booked** to **book** and **booking** to **book**
UN*;ED;ING will join **undo** to **do, booked** to **book** and **booking** to **book**
BLOODY will join **absobloodylutely** to **absolutely**

The process can be left to run quickly and automatically, or you can have it confirm with you before joining each one. Automatic lemmatisation, like search-and-replace spell-checking, can produce oddities if just left to run!
To stop in the middle of auto-joining, press Escape.

Tip

With a previously saved list, try auto-joining *without* confirming the changes (or choose *Yes to All* during it). Then choose the Alphabetical (as opposed to Frequency) version of the list and sort on Lemmas (by pressing the *Lemmas* column heading). You will see all the joined entries at the top of the list. It may be easier to Unjoin 322 (Ctrl+F4) any mistakes than to confirm each one... Finally, sort on the *Word* and save.

See also: Lemmatisation 322

9.7.4 choosing lemma file

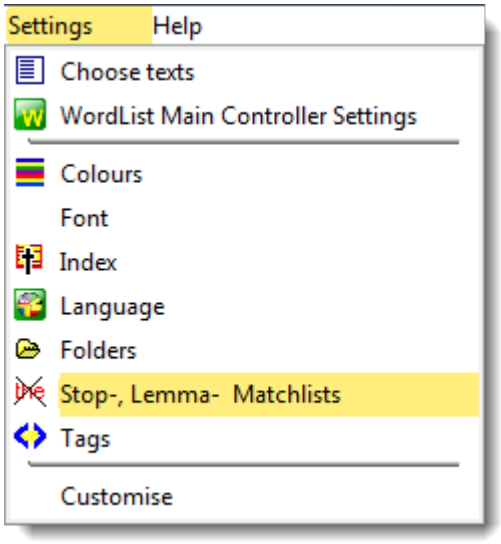
The point of it...

You may choose to lemmatise all items in the current word-list using a standard text file which groups words which belong together (**be -> was, is, were**, etc.). It will be very useful if you want to lemmatise lots of word lists, and is much less "hit-and-miss" than auto-joining 326 using a template.

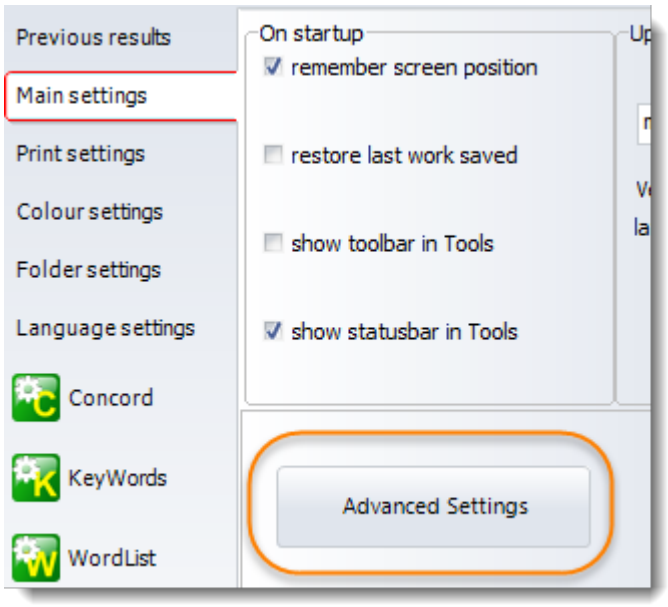
There are English-language lemma lists at https://lexically.net/wordsmith/support/lemma_lists.html. And you can build your own easily 444 if you have a corpus with lemma mark-up, for example the BNC.

How to do it

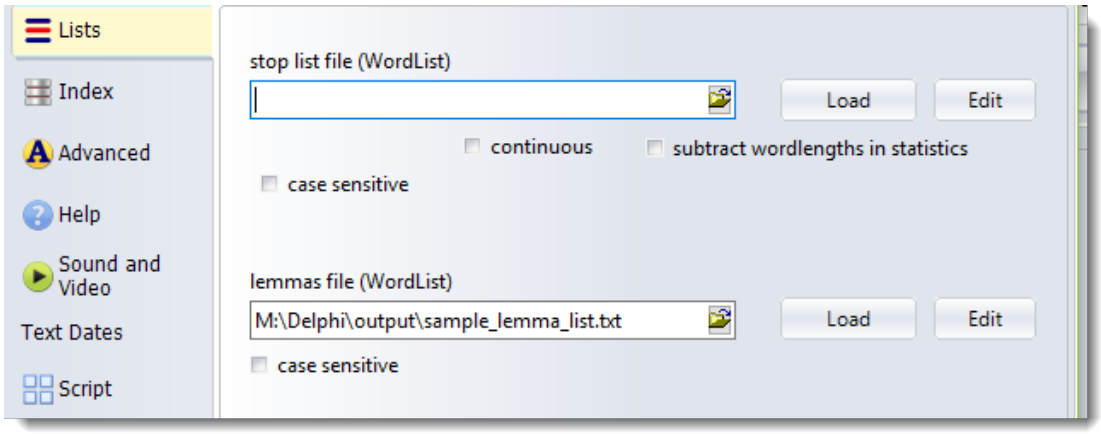
Lemma list settings are accessed via the Lists option in the WordList menu



or an Advanced Settings button in the Controller



followed by



Choose the appropriate button (for Concord, KeyWords or WordList) and type the file name or browse for it, then Load it.

The file should contain a plain text list of lemmas with items like this:

```
BE -> AM, ARE, WAS, WERE, IS
GO -> GOES, GOING, GONE, WENT
```

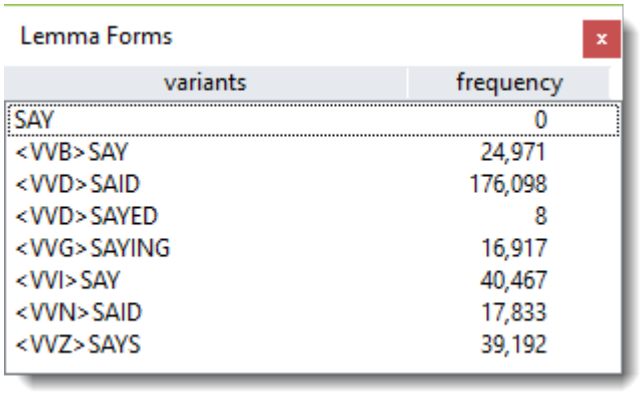
WordSmith then reads the file and displays them (or a sample if the list is long). The format allows any alphabetic or numerical characters in the language the list is for, plus the single apostrophe, space, underscore. In other words, if you mistakenly put **GO = GOES** that line won't be included because of the = symbol.

A lemma list can be *case sensitive*: e.g. if using with a case sensitive word list 371.

The actual processing of the list will take place when you compute your word list, key word list or concordance or when you choose the menu option *Match Lemmas* ( ) in WordList, Concord or KeyWords. See Match List 82 for a more detailed explanation, with screenshots. Lemmatising occurs before any stop list 141 is processed.

What if my text files don't contain the headword of the lemma?

Suppose you are matching **AM, ARE** etc with **BE** as in the list above, but your texts don't actually contain the word **BE**. In that case the tool will insert **BE** with zero frequency and add **AM, ARE** etc as needed.



In this screen-shot the form **SAY** is shown with zero frequency. That's because all the data had **<VV>** mark-up. No case of **SAY** without mark-up was present but the lemma file specified **SAY** as the head word, so a new entry showing the head word gets created.

See also: Lemmatisation 322, Building a lemma file 444, Match List 82, Stop List 141, Lemmatisation in Concord 211

9.8 WordList Index

9.8.1 what is an Index for?

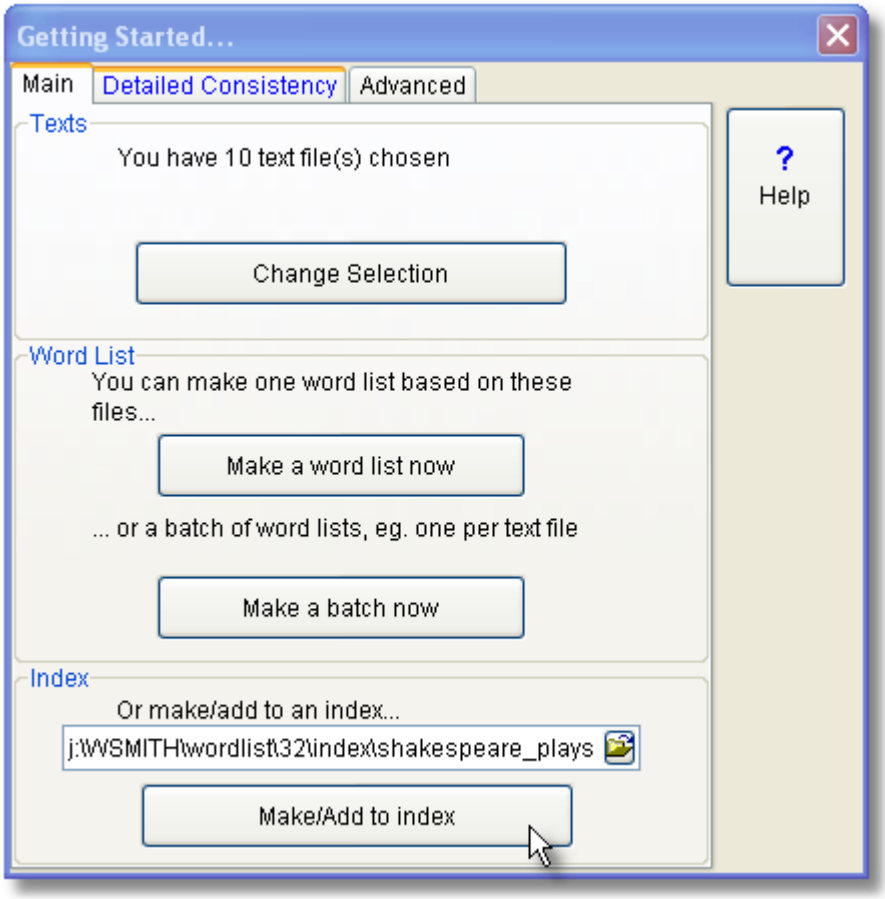
the point of it

1. One of the uses for an Index is to record the positions of all the words in your text file, so that you can subsequently see which word came in which part of each text. Another is to speed up access to these words, for example in concordancing. If you select one or more words in the index and press , you get a speedy concordance.
2. Another is to compute "Mutual Information" or other statistical 342 scores which relate word types to each other.
3. Or you can use an index to see word clusters 332.
4. Finally, an index is needed to generate concgram 13 searches.

See also Making an Index List 330, Viewing Index Lists 338, Exporting index data 340, Index settings 366, find filenames for word clusters 320, WordList Help Contents 305, WSConcgram 13

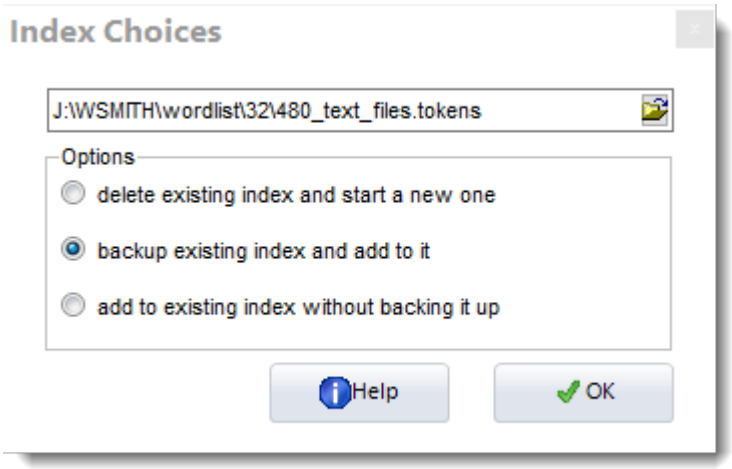
9.8.2 making a WordList Index

The process is just like the one for making a word-list except that after choosing your texts and ensuring you like the index filename, you choose the bottom button here:

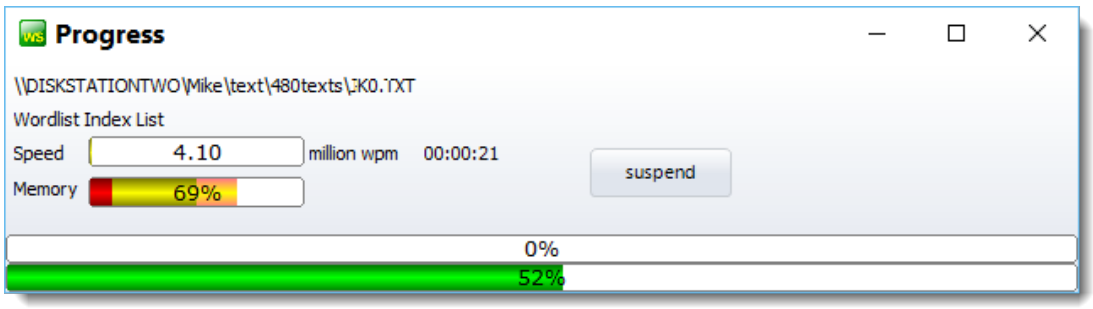


In this screenshot above, the basic filename is **shakespeare_plays**: WordSmith will add **.tokens** and **.types** to this basic filename as it works. Two files are created for each index:
.tokens file: a large file containing information about the position of every word token in your text files.
.types file: knows the individual word types.

If you choose an existing basic filename which you have already used, **WordList** will check whether you want to add to it or start it afresh:



An index permits the computation of word clusters 332 and statistical relationships 342 scores for each word type. The screenshot below shows the progress bars for an index; here it worked out at a rate of about 4 million words per minute.



adding to an index

To add to an existing index, just choose some more texts and choose *File* | *New* | *Index*. If the existing file-name is already in use for an index, you will be asked whether to add more or start it afresh as shown above.

See also Using Index Lists 330, Viewing Index Lists 338, WordList Help Contents 305.

9.8.3 index clusters

WordList clusters

A word list doesn't need to be of single words. You can ask for a word list consisting of two, three, up to eight words on each line.

Example

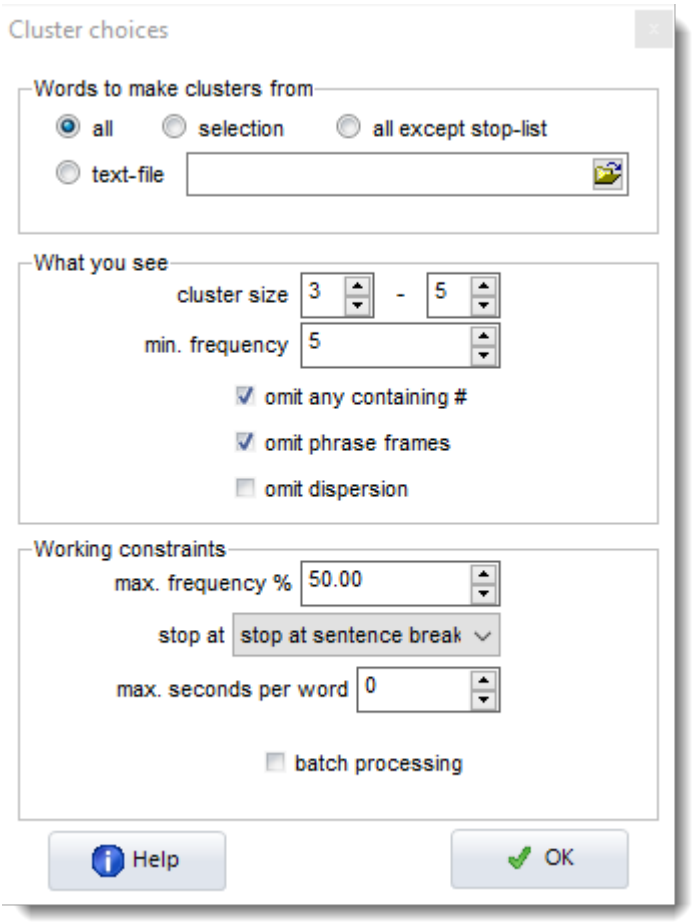
Here is a small set of 3-4word clusters involving the word **money**.

N	Word	Freq.	%	Texts	%	Dispersion	Le...s	Set
1	LOT OF MONEY	37		22	8.15%	0.00		3
2	A LOT OF MONEY	32		19	7.04%	0.00		4
3	A LOT OF	32		19	7.04%	0.00		3
4	OF THE MONEY	29		20	7.41%	0.00		3
5	MONEY OUT OF	23		17	6.30%	0.00		3
6	THE MONEY TO	18		17	6.30%	0.00		3
7	MONEY IN THE	17		15	5.56%	0.00		3
8	AMOUNT OF MONEY	16		14	5.19%	0.00		3
9	VALUE FOR MONEY	15		10	3.70%	0.00		3
10	SUMS OF MONEY	15		13	4.81%	0.00		3

Some of them are plausible multi-word units. The number of words in each cluster gets inserted in the Set column 43.

How to do it

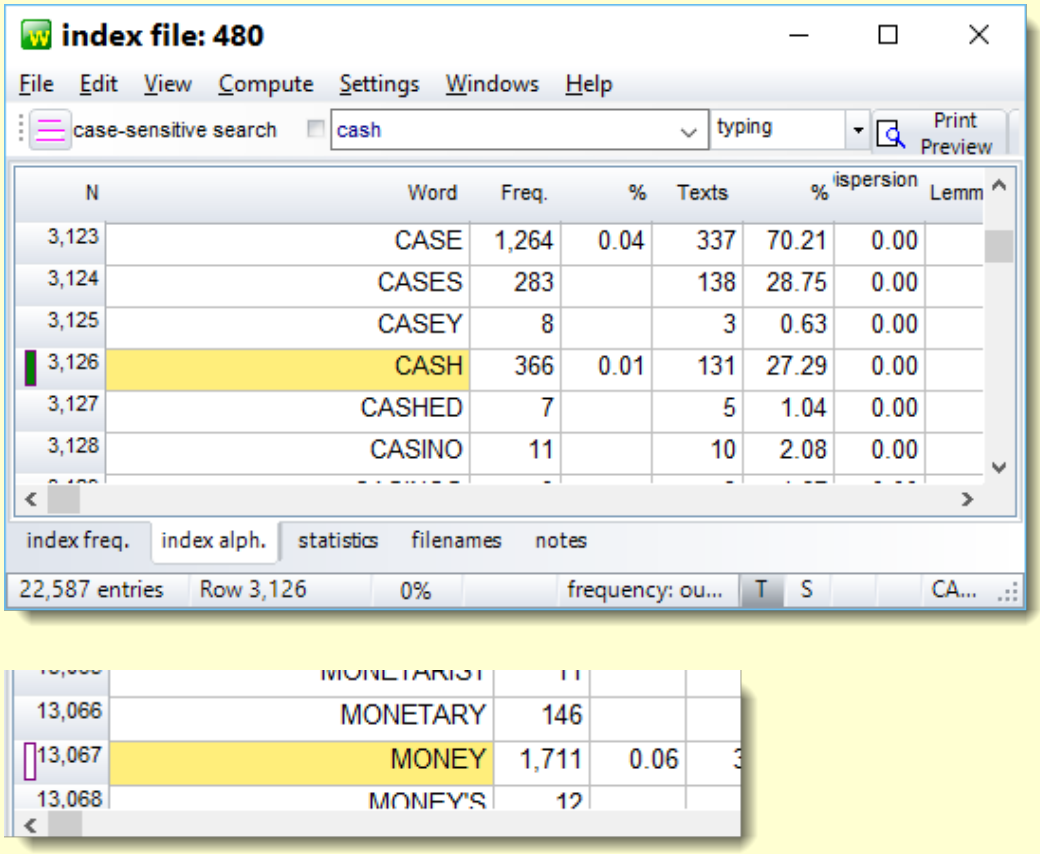
To do cluster 556 processing in WordList, first make an index 330.
Then open 338 the index. Now choose *Compute* | *Clusters*.



Words to make clusters from

- "all" : all the clusters involving all words above a certain frequency (this will be slow for a big corpus like the BNC), or
- "selection": clusters only for words you've selected (e.g. 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 480 texts, I selected **`cash`** and then **`money`** by clicking numbers 3,126 and 13,067.



The process will take time. .

What you see

The *cluster size* must be between 2 and 8 words.
The *min. frequency* is the minimum number of each that you want to see.
omit #: if selected, this won't show any clusters involving numbers and dates
omit phrase frames: see phrase frames 336 section below.
omit dispersion: if checked does not compute dispersion 354.
Here the user has chosen to see any 3-5-word clusters that appear 5 or more times.

Working constraints

The "max. frequency %" setting is to speed the process up.

in more detail...

It means the maximum frequency percentage which the calculation of clusters for a given word will process. This is because there are lots and lots of the very high frequency items and you may well not be interested in clusters which *begin* with them. For example, the item **the** is likely to be about 6% of any word-list (about 6 million of them in the BNC therefore), and you might not want clusters starting **the...** -- if so, you might set the max. percent to 0.5% or 0.1% (which for the BNC corpus will cut out the top 102 frequency words). You will still get clusters which include very high frequency items in the middle or end, like the **a** in **book a table**, but would not get **in my book**, which begins with the very high frequency word **in**. The more words you include, the longer the process will take....

Stop at, like Concord clusters 203, offers a number of constraints, such as sentence and other punctuation-marked breaks 217. The idea is that a 5-word cluster which starts in one sentence and continues in the next is not likely to make much sense.

Max. seconds per word is another way of controlling how long the process will take. The default (0) means no limit. But if you set this e.g. to 30 then as WordList processes the words in order, as soon as one has taken 30 seconds no further clusters will be collected starting with that word.

Omit dispersion will save a lot of memory and time. That's because by default each cluster uses a whole extra lot of memory, keeping track of where each occurrence comes in the corpus.

batch processing allows you to create a whole set of cluster word-lists at one time.

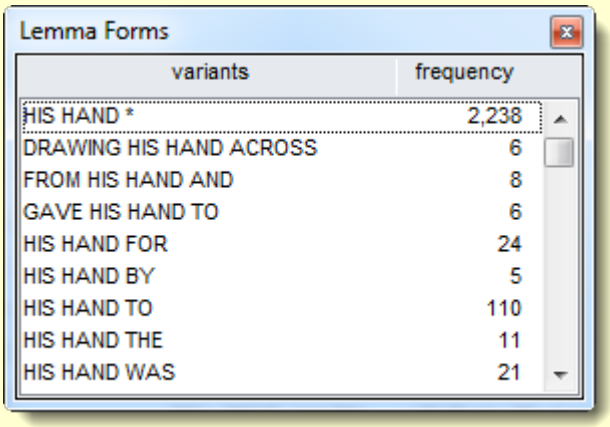
Phrase frames

These are what William H. Fletcher has defined as *phrase-frames*, i.e. "groups of wordgrams identical but for a single word", in his kfNgram program.

Here, processing 23 Dickens novels shows lots of phrase frames where the wildcard word is represented with *.

WHICH HE *	2,280	0.05	0	0.(	which he *[2280] at which he had[16] by which he was
IT * TO	2,279	0.05	0	0.(	it * to[2279] and it appeared to[5] and it seemed to[8]
SHE WAS *	2,249	0.05	0	0.(	she was *[2249] and she was not[6] and she was very
HIS HAND *	2,238	0.05	0	0.(	his hand *[2238] drawing his hand across[6] from his h
A MAN *	2,236	0.05	0	0.(	a man *[2236] a man for[7] a man to[63] a man the[8]
I * IT	2,223	0.05	0	0.(	i * it[2223] and i hope it[8] and i thought it[10] and i fou

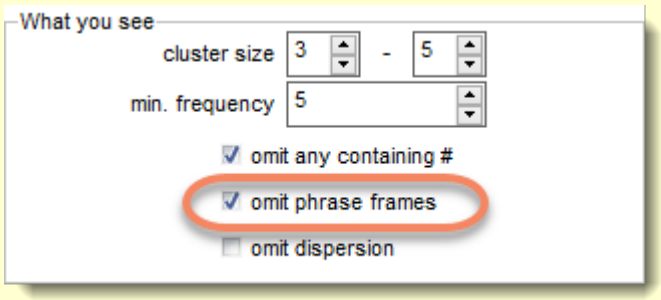
If you double-click the lemmas column (highlighted here in yellow), you get to see the detail.



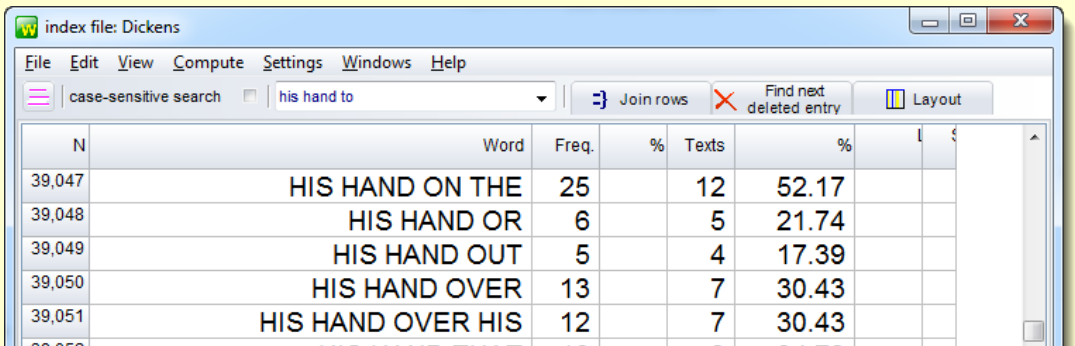
The process joins all the variants of the phrase in the Lemmas column. In the word list itself they will appear deleted (because they have been joined to another item, the phrase frame). You can un-join them all if you want (*Edit* | *Joining* | *Unjoin* or *Unjoin all*).

Omit phrase frames?

If you don't want to see phrase frames, select the *omit phrase frames* option.



Here below, the listing has all **`his hand`** sequences together but not **`drawing his hand across`**, **`gave his hand to`**, etc. as shown in the phrase frame view above.



It's a word list

Finally, remember this listing is just like a single-word word list. You can save it as a **.lst** file and open it again at any time, separately from the index.

See also: clusters in single-word list 306, find the files for specific clusters 320, clusters in Concord 556

9.8.4 join clusters

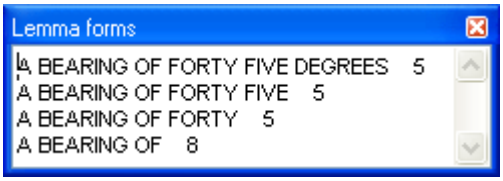
The idea is to group clusters like

I DON'T THINK
NO I DON'T THINK
I DON'T THINK SO
I DON'T THINK THAT
etc.

You can join them up in a process like lemmatisation 322, either so that the smaller clusters get merged as 'lemmas' of a bigger one, or so that the smaller ones end up as 'lemmas'.

In this screenshot, shorter clusters have been merged with longer ones so that **A BEARING OF FORTY-FIVE DEGREES** relates to several related clusters:

visible by double-clicking the lemmas to show something like this:



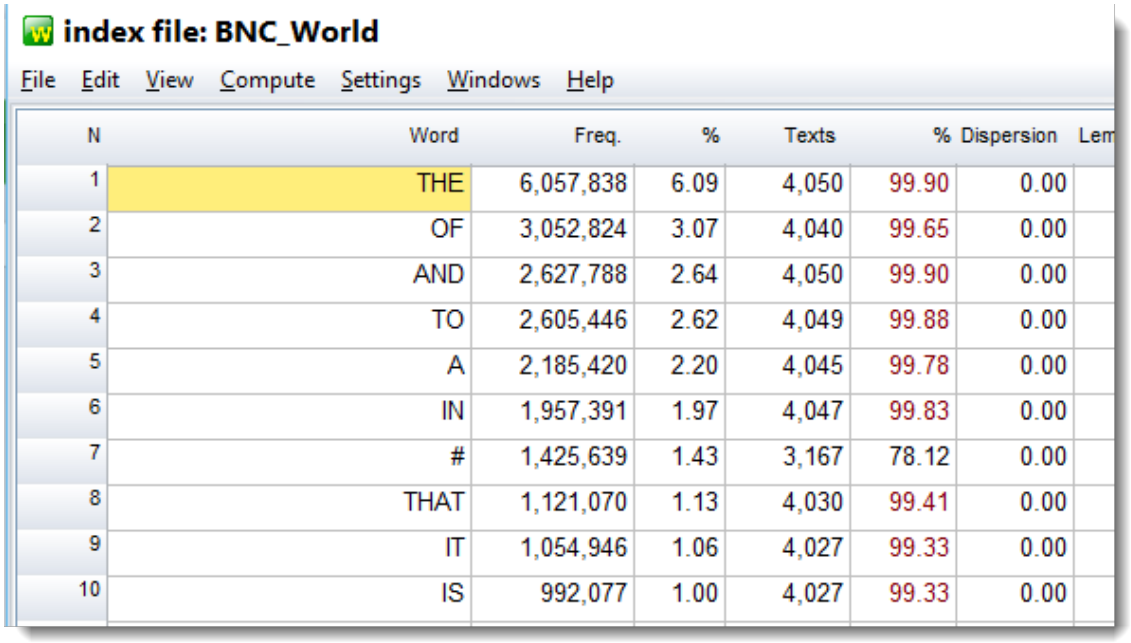
How to do it

Choose *Edit* | *Join* | *Join Clusters* in the WordList menu. The process takes quite a time because each cluster has to be compared with all those in the rest of the list; interrupt it if necessary by pressing Suspend 144.

9.8.5 index lists: viewing

In WordList, open an index as you would any other kind of word-list file -- using File | Open. The filename will end **.tokens.** Easier, in the *Controller* | *Previous lists*, choose any index you've made and double-click it.

The index *looks* almost exactly like a large word-list. (Underneath, it "knows" a lot more and can do more but it looks the same.)



N	Word	Freq.	%	Texts	%	Dispersion	Lem
1	THE	6,057,838	6.09	4,050	99.90	0.00	
2	OF	3,052,824	3.07	4,040	99.65	0.00	
3	AND	2,627,788	2.64	4,050	99.90	0.00	
4	TO	2,605,446	2.62	4,049	99.88	0.00	
5	A	2,185,420	2.20	4,045	99.78	0.00	
6	IN	1,957,391	1.97	4,047	99.83	0.00	
7	#	1,425,639	1.43	3,167	78.12	0.00	
8	THAT	1,121,070	1.13	4,030	99.41	0.00	
9	IT	1,054,946	1.06	4,027	99.33	0.00	
10	IS	992,077	1.00	4,027	99.33	0.00	

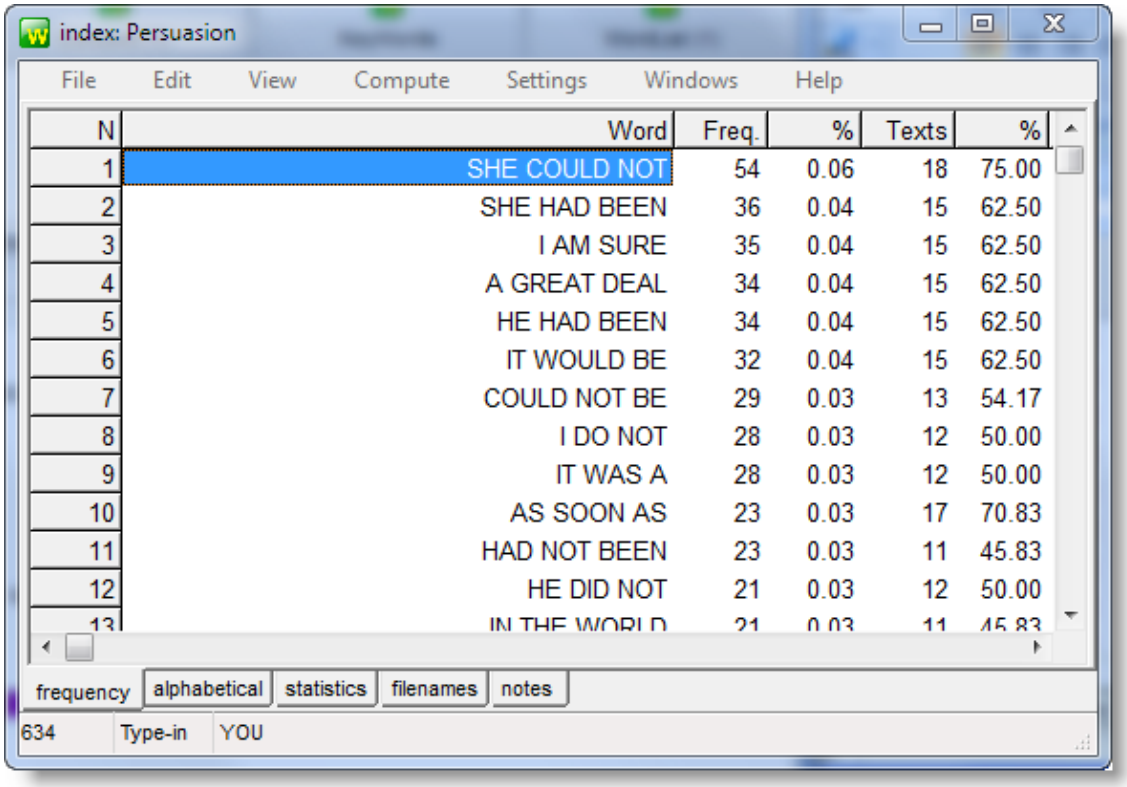
The picture above shows the top 10 words in the BNC Corpus. Number 7 (#) represents numbers or words which contain numbers such as £50.00. These very frequent words are also very consistent -- they appear in nearly 80% of the 4,054 texts of BNC.

In the view below, you see words sorted by the number of Texts: all these words appeared in 10 texts in the corpus but their frequencies vary.

63,936	RESUBMIT	14		10	0.25	0.00
63,937	RESTOCKED	10		10	0.25	0.00
63,938	RESISTIVE	33		10	0.25	0.00
63,939	RESISTIBLE	10		10	0.25	0.00
63,940	REREDOS	14		10	0.25	0.00
63,941	REPROACHING	10		10	0.25	0.00
63,942	REPRESSIVELY	13		10	0.25	0.00
63,943	REPRESSES	11		10	0.25	0.00
63,944	REPOPULATE	11		10	0.25	0.00
63,945	REPLICABLE	15		10	0.25	0.00

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

But its best use, to start with, is to generate word clusters 332 like these:



See also Making an Index List 330, WordList clusters 332, WordList Help Contents 305.

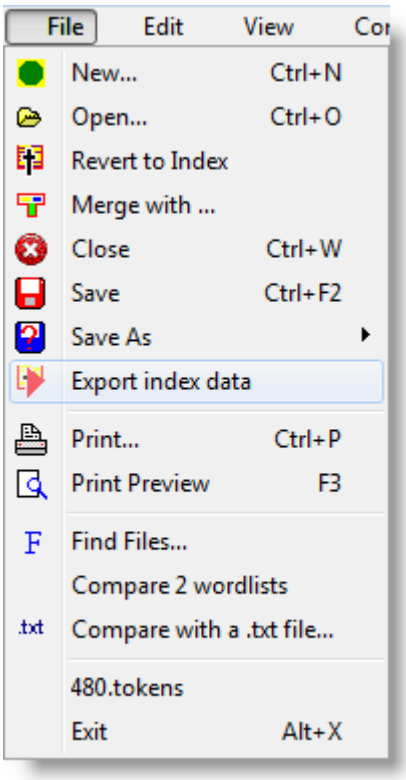
9.8.6 index exporting

The point of it...

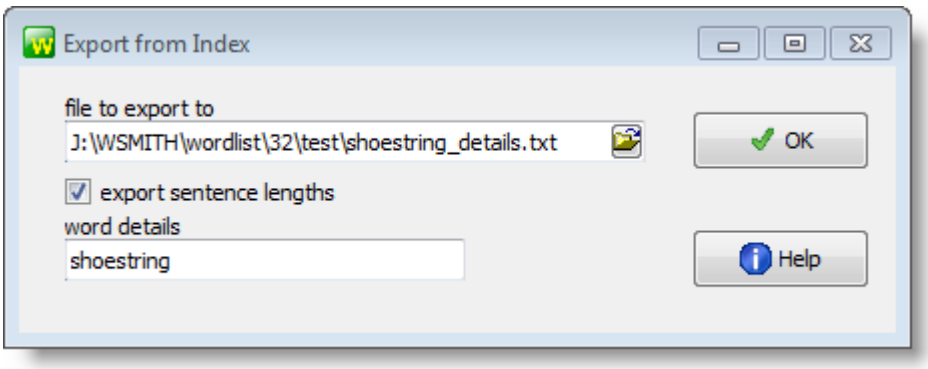
An index file knows the position of every single word in your corpus and it is possible therefore to ask it to supply specific data. For example, the lengths of each sentence or each text in the corpus (in words), or the position of each occurrence of a given word.

How to do it

With an index open, choose File | Export index data,

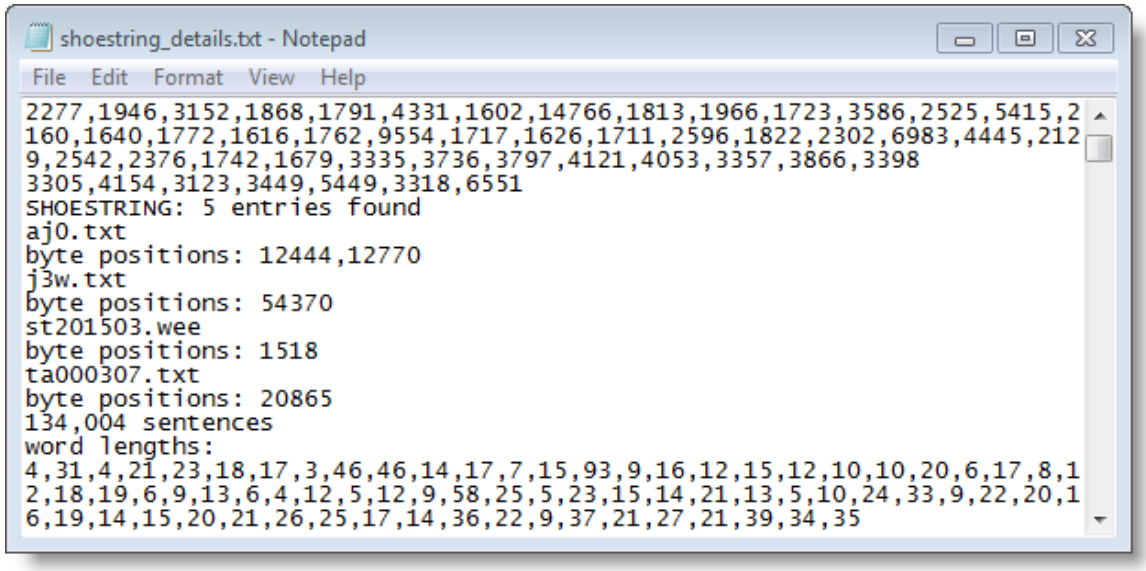


then complete the form with what you need.



Here we have chosen to export the details about the word **SHOESTRING** in a given index, and to get to see all the sentence lengths (of all sentences in the corpus, not just the ones containing that word).

A fragment of the results are shown here:



At the top there are word-lengths of some of the 480 text files, the last of which was 6551 words long; then we see the details of 5 cases of the word **SHOESTRING** in the corpus, which appeared twice in text AJ0.txt, once in J3W.txt etc.; finally we get the word-lengths of all the sentences in the corpus : the first one only 4 words long.

This process will be quite slow if you request a lot of data. If you don't check the sentence lengths you will still get text lengths; it wil be quicker if you leave the word details space empty.

9.9 menu search

Using the menu you can search for a sub-string within an entry -- e.g. all words containing "fore" (by entering ***fore*** -- the asterisk means that the item can be found in the middle of a word, so ***fore** will find *before* but not *beforehand*, while ***fore*** will find them both). These searches can be repeated.
This function enables you to find parts of words so that you can edit your word-list, e.g. by joining two words as one.
You can search for ends or middles of words by using the * wildcard.
Thus ***TH*** will find *other, something*, etc.
***TH** will find *booth, sooth*, etc.
You can then use **F8** to repeat your last search.

The search hot keys are:

F8 repeat last search (use in conjunction with F10 or F11)

F10	search forwards from the current line
F11	search backwards from the current line
F12	search starting from the beginning

This function is handy for lemmatization 322 (joining words which belong under one entry, such as *seem/ seems/ seemed/ seeming* etc.)

See also: searching for an entry by typing 106

9.10 relationships between words

9.10.1 mutual information and other relations

the point of it

A Mutual Information (MI) score relates one word to another. For example, if *problem* is often found with *solve*, they may have a high mutual information score. Usually, *the* will be found much more often near *problem* than *solve*, so the procedure for calculating Mutual Information takes into account not just the most frequent words found near the word in question, but also whether each word is often found elsewhere, well away from the word in question. Since *the* is found very often indeed far away from *problem*, it will not tend to be related, that is, it will get a low MI score.

There are several other alternative statistics: you can see examples of how they differ here 342.

This relationship is nearly always bi-lateral (except for Conditional probability and Delta probability): in the case of the cluster *of course*, most statistics don't distinguish between the good likelihood of finding **`of`** near **`course`**, and the much lower likelihood of finding **`course`** near **`of`**. Conditional probability and Delta probability distinguish between the two frequencies, measuring the frequency of one in relation to the joint frequency score. See collocation relationship 207 for an example.

There are various different formulae 540 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 524, 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 Relationships Display 342 and Collocation relationship 207.

Settings

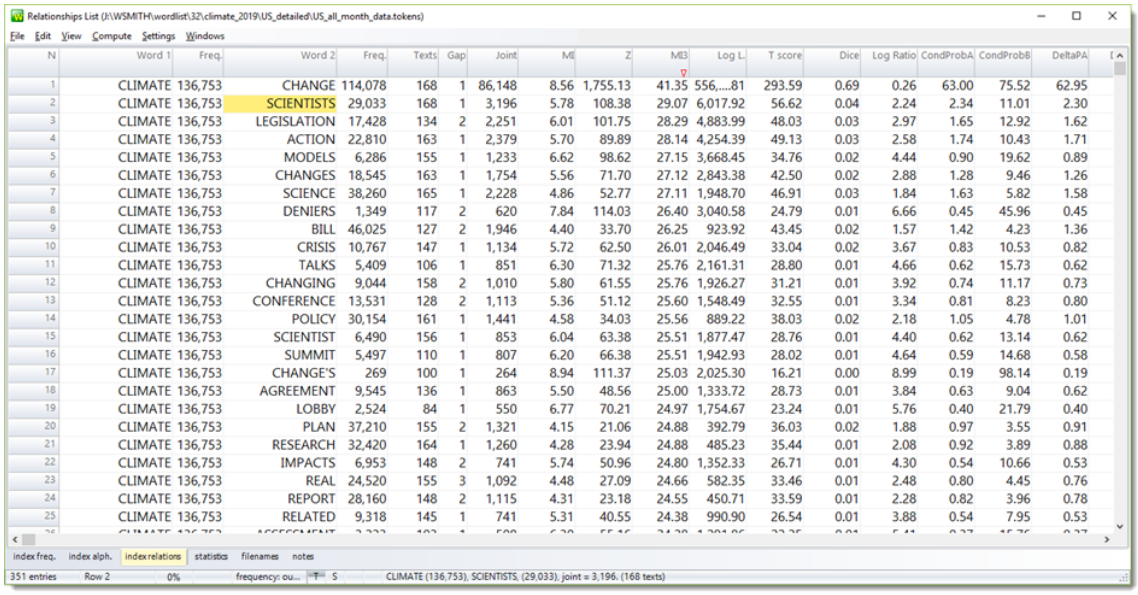
The Relationships settings are found in the Controller 5 under *Main Settings* | *Advanced* | *Index* 366 or in a menu option in **WordList**.

See also: Relationships Display 342, Computing Relationships 344, Formulae 540, Making an Index List 330, Viewing Index Lists 338, WordList Help Contents 305.

See Oakes 524 for further information about Mutual Information, Dice, MI3 etc.

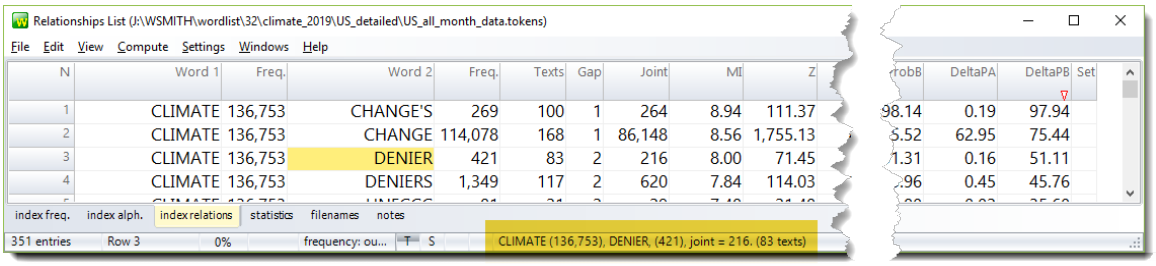
9.10.2 relationships display

The Relationships procedure contains a number of columns and uses various formulae 540:



Word 1: the first word in a pair, followed by Freq. (its frequency in the whole index).
Word 2: the other word in that pair, followed by Freq. (its frequency in the whole index). If you have computed "to right only 346", then Word 1 precedes Word 2.
Texts: the number of texts this pair was found in (there were 23 in the whole index).
Gap: the most typical distance between Word 1 and Word 2.
Joint: their joint frequency over the entire span 344 (not just the joint frequency at the typical gap distance).

In line 2 of this display, **SCIENTISTS** occurs over 29,000 times in the whole index (based on US climate change news text), and it comes joint with **CLIMATE** over 3,000 times. This is shown in the status bar. The gap is 1 because scientists, in these data, typically comes 1 words away from climate. The pair **climate * scientists** comes in 168 texts. (This search was computed using the to right only 346 setting mentioned above).



There is a big difference in the DeltaPA and DeltaPB scores for the pair **climate * denier**. That's because roughly 216 in 421 is a much bigger proportion than 216 in 136,000. Therefore in these data, **denier** is a better predictor of **climate** than **climate** is of **denier**.

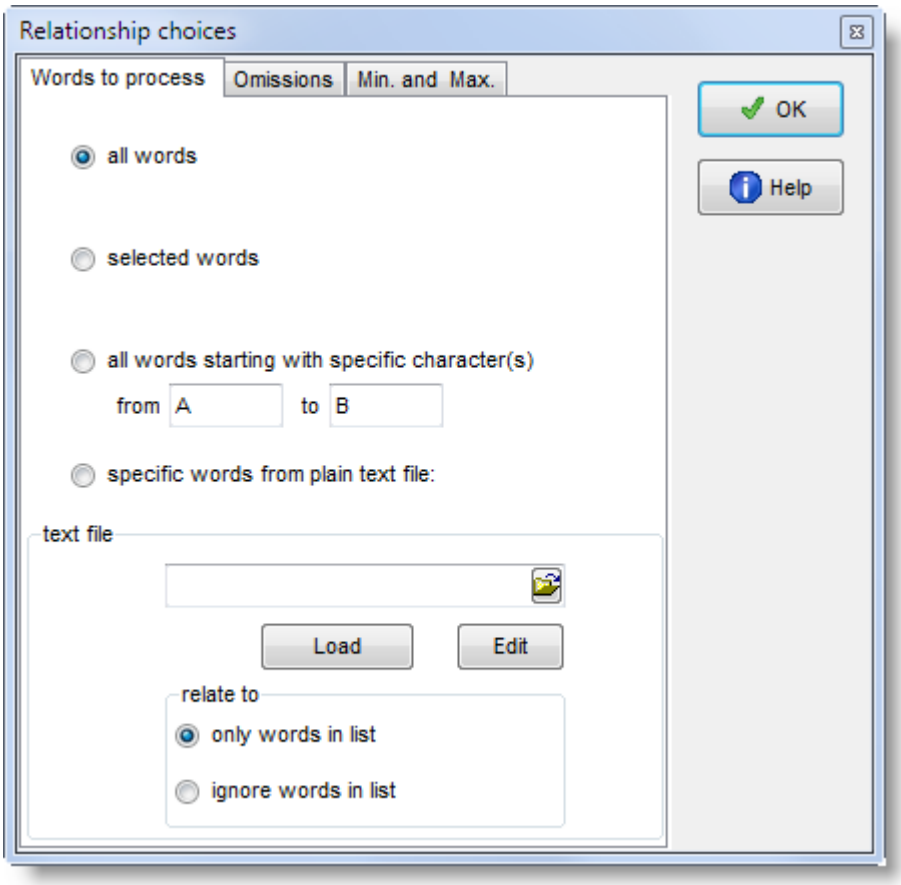
As usual, the data can be sorted by clicking on the headers.

See also: Formulae 540, Mutual Information and other relationships 342, Computing Relationships 344, Making an Index List 330, Viewing Index Lists 338, WordList Help Contents 305.

See Oakes 524 for further information about the various statistics offered.

9.10.3 relationships computing

To compute these relationship statistics you need a WordList Index 330. Then in its menu, choose Compute | Relationships.



words to process

You can choose whether to compute the statistics for all entries, or only any selected (highlighted) entries, or only those between two initial characters e.g. between A and D, or indeed to use your own specified words only.

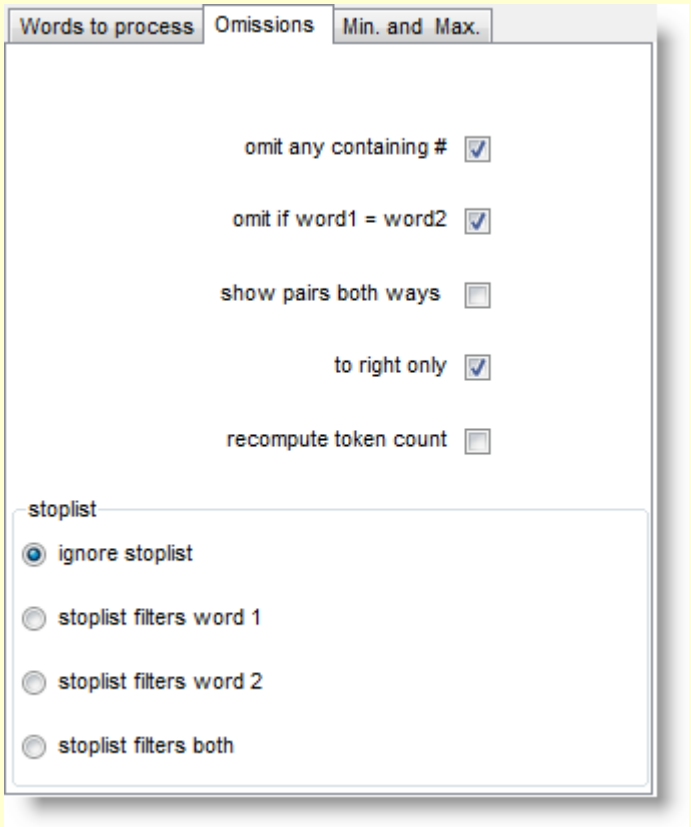
If you wish to select only a few items for MI calculation, you can mark them first 107 (with). Or you can always do part of the list (eg. A to D) and later merge 310 your mutual-information list with another (E to H).

Text file

Alternatively you may choose to use only items from a plain text file constructed using the same syntax as a match-list file., or to use all items except ones from your plain text file.

omissions

omit any containing # will cut out numbers, and *omit if word1=word2* is there because you might find that **GOOD** is related to **GOOD** if there are lots of cases where these 2 are found near each other.



show pairs both ways allows you to locate all the pairs more easily because it doubles up the list. For example, suppose we have a pair of words such as **HEAVEN** and **EARTH**. This will normally enter the list only in one order, let us say **HEAVEN** as word 1 and **EARTH** as word 2. If you're looking at all the words in the Word 1 column, you will not find **EARTH**. If you want to be able to see the pair as both **HEAVEN** - **EARTH** and **EARTH** - **HEAVEN**, select *show pairs both ways*. Here we can see this with **DUST** and **WITH**

N	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI
1	DUST	59	WITH	7,148	9	1	11	4.41
2	WITH	7,148	DUST	59	9	1	11	4.41

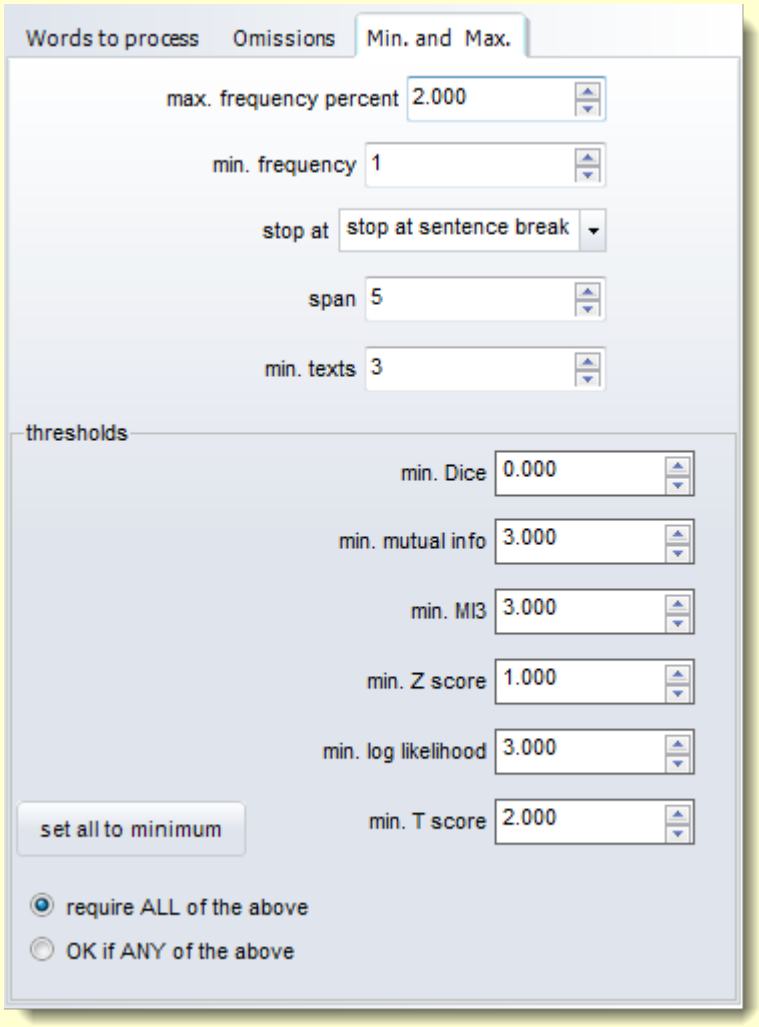
to right only: if this is checked, possible relations are computed to the right of the node only. That is, when considering **DUST**, say, cases of **WITH** to the right will be noticed but cases where **WITH** is to the left of **DUST** would get ignored.

N	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI	Z
1	DUST	59	WITH	7,148	5	1	6	3.54	0.37

Here, the number of texts goes down to 5 from 9, MI score is lower, etc, because the process looks only to the right. (In the case of a right-to-left language like Arabic, the processing is still of the words following the node word.)

recompute token count allows you to get the number of tokens counted again e.g. after items have been edited or deleted.

min. and max



max. frequency percent: ignores any tokens which are more frequent than the percentage indicated. Set the maximum frequency, for example, to 0.5% to cut out words whose frequency is greater than that.(The point of this is to avoid computing mutual information for words like **the** and **of**, which are likely to have a frequency greater than say 1.0%. For example 0.5%, in the case of the BNC, would mean ignoring about 20 of the top frequency words, such as **WITH, HE, YOU**. 0.1% would cut about 100 words including **GET, BACK, BECAUSE**. If you want to include all words, then set this to 100.000)

min. frequency: the minimum frequency for any item to be considered for the calculation. (Default = 5; a minimum frequency of 5 means that no word of frequency 4 or less in the index will be visible in the relationship results. If an item occurs only once or twice, the relationship is unlikely to be informative.)

stop at allows you to ignore potential relationships e.g. across sentence boundaries. It has to do with whether breaks such as punctuation or sentence breaks[217] determine that one word cannot be related to another. With stop at sentence break, "**I wrote the letter. Then I posted it**" would not consider **posted** as a possible collocate of **letter** because there's a sentence break between them.

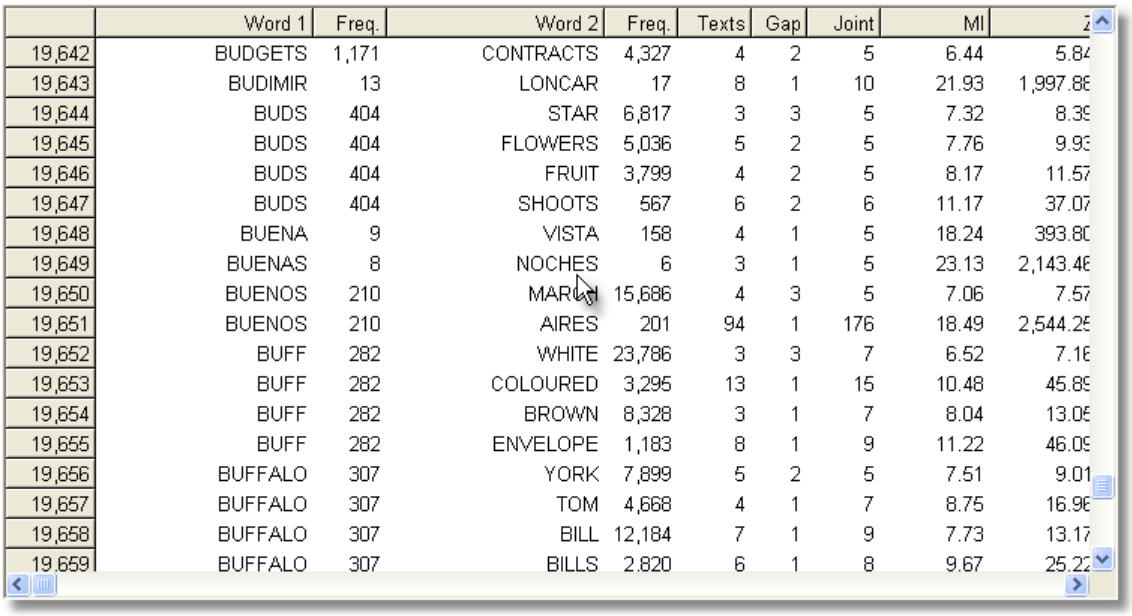
span: the number of intervening words between collocate and node. With a span of 5, the node **wrote** would consider **the, letter, then, I** and **posted** as possible collocates if *stop at* were set at *no limits* in the example above.

min. texts: the minimum number of texts any item must be found in to be considered for the calculation.

min. Dice/mutual info.MI3 etc: the minimum number which the MI or other selected statistic must come up with to be reported. A useful limit for MI is 3.0. Below this, the linkage between node and collocate is likely to be rather tenuous.

Choose whether ALL the values set here are used when deciding whether to show a possible relationship or ANY. (Each threshold can be set between -9999.0 and 9999.0.)

Computing the MI score for each and every entry in an index takes a long time. Don't forget to save your results afterwards!



	Word 1	Freq.	Word 2	Freq.	Texts	Gap	Joint	MI	
19,642	BUDGETS	1,171	CONTRACTS	4,327	4	2	5	6.44	5.84
19,643	BUDIMIR	13	LONCAR	17	8	1	10	21.93	1,997.88
19,644	BUDS	404	STAR	6,817	3	3	5	7.32	8.39
19,645	BUDS	404	FLOWERS	5,036	5	2	5	7.76	9.93
19,646	BUDS	404	FRUIT	3,799	4	2	5	8.17	11.57
19,647	BUDS	404	SHOOTS	567	6	2	6	11.17	37.07
19,648	BUENA	9	VISTA	158	4	1	5	18.24	393.80
19,649	BUENAS	8	NOCHES	6	3	1	5	23.13	2,143.48
19,650	BUENOS	210	MAR	15,686	4	3	5	7.06	7.57
19,651	BUENOS	210	AIRES	201	94	1	176	18.49	2,544.25
19,652	BUFF	282	WHITE	23,786	3	3	7	6.52	7.18
19,653	BUFF	282	COLOURED	3,295	13	1	15	10.48	45.89
19,654	BUFF	282	BROWN	8,328	3	1	7	8.04	13.05
19,655	BUFF	282	ENVELOPE	1,183	8	1	9	11.22	46.09
19,656	BUFFALO	307	YORK	7,899	5	2	5	7.51	9.01
19,657	BUFFALO	307	TOM	4,668	4	1	7	8.75	16.96
19,658	BUFFALO	307	BILL	12,184	7	1	9	7.73	13.17
19,659	BUFFALO	307	BILLS	2,820	6	1	8	9.67	25.22

See also Collocates 206, Mutual Information Settings 342, Mutual Information Display 342, Detailed Consistency Relations 319, Making an Index List 330, Viewing Index Lists 338, Recompute Token Count 349, WordList Help Contents 305.

9.11 recompute tokens

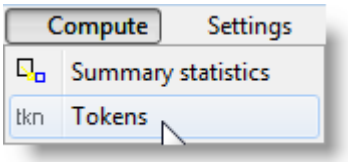
Why recompute the tokens?

To compute relations such as Mutual Information 342 or Keyness 287 we need an estimate of the total number of running words (let's call it TNR) in the text corpus from which the data came. It is tricky to decide what actually counts as the TNR. Not only are there problems to do with hyphenation 146, apostrophes and other non-letters 146 in the middle of a word, numbers 146, words cut out because of a stoplist 141 etc, but also a decision whether TNR should in principle include all of those or in principle include only the words or clusters now in the list in question. In practice for single-word word lists this usually makes little difference. In the case of word clusters, however, there might be a big difference between the TNR words and TNR clusters, and anyway what exactly is meant by running clusters of words if you think about how they are computed 556?

For most normal purposes, the total number of running words (tokens 533) computed when the word list or index was created will be used for these statistical calculations.

How to do it

Compute | Tokens



What it affects

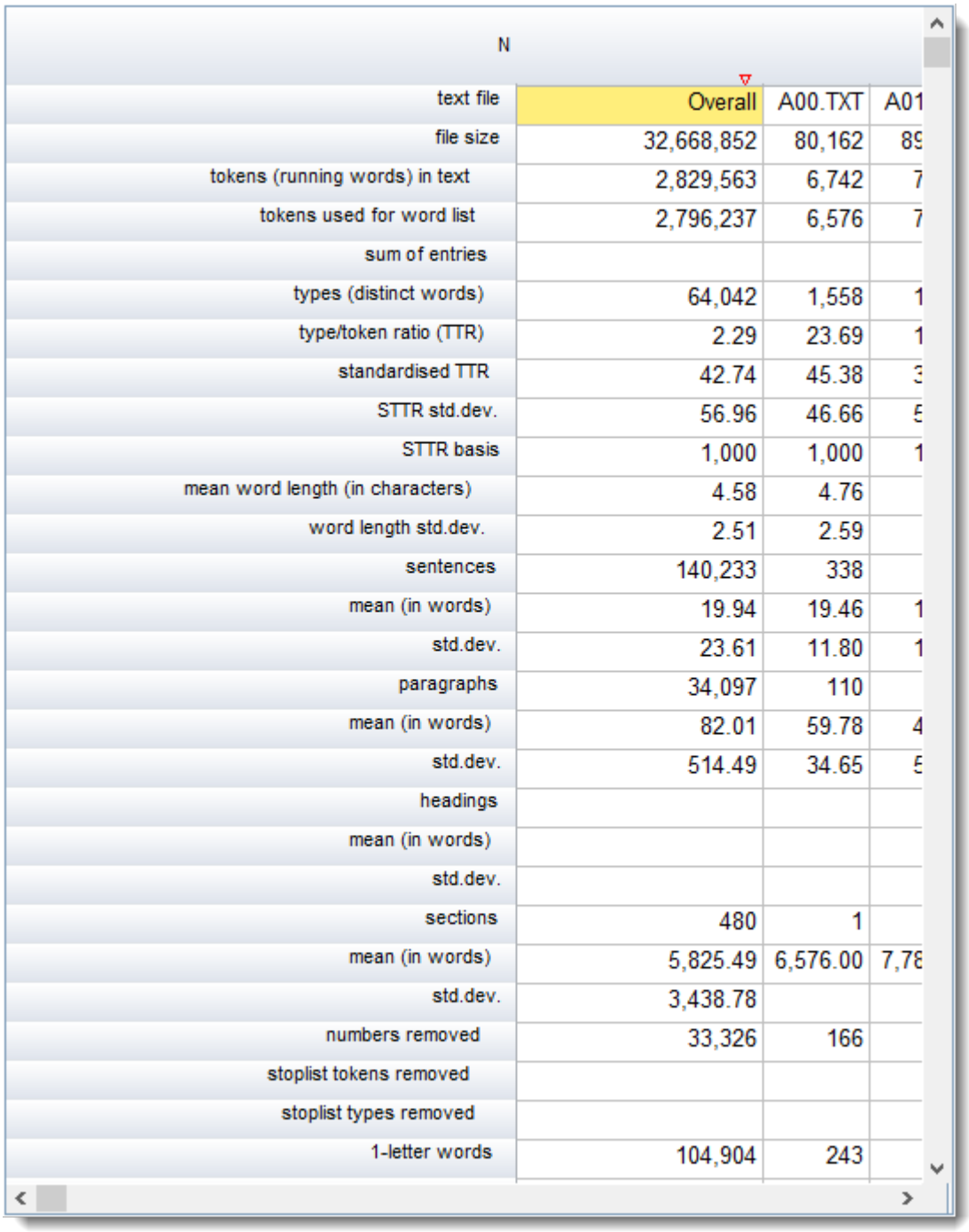
Any decision made here will apply equally both to the node and the collocate whether these are clusters or single words, or to the little word-list and the reference corpus word-list in the case of key words calculations.

If you do choose to recompute the token count, then the TNR will be calculated as the total of the word or cluster frequencies for those entries still left in the list. After any have been zapped or if a minimum frequency above 1 is used the difference may be quite large.

If you choose *not* to recompute, the total number of running words (tokens) computed when the word list or index was created will be used.

9.12 statistics

Visible by clicking the Statistics tab at the bottom of a **WordList** window:



N			
text file	Overall	A00.TXT	A01
file size	32,668,852	80,162	89
tokens (running words) in text	2,829,563	6,742	7
tokens used for word list	2,796,237	6,576	7
sum of entries			
types (distinct words)	64,042	1,558	1
type/token ratio (TTR)	2.29	23.69	1
standardised TTR	42.74	45.38	3
STTR std.dev.	56.96	46.66	5
STTR basis	1,000	1,000	1
mean word length (in characters)	4.58	4.76	
word length std.dev.	2.51	2.59	
sentences	140,233	338	
mean (in words)	19.94	19.46	1
std.dev.	23.61	11.80	1
paragraphs	34,097	110	
mean (in words)	82.01	59.78	4
std.dev.	514.49	34.65	5
headings			
mean (in words)			
std.dev.			
sections	480	1	
mean (in words)	5,825.49	6,576.00	7,78
std.dev.	3,438.78		
numbers removed	33,326	166	
stoplist tokens removed			
stoplist types removed			
1-letter words	104,904	243	

Overall results take up the first column. Details for the individual text files follow to the right.

Statistics include:

number of files involved in the word-list
file size (in bytes, i.e. characters)
running words in the text (***tokens***)
tokens used in the list (would be affected by using a stoplist 141 or changes to minimum settings 370)

sum of entries: choose *Compute* | *Tokens* to see, otherwise this will be blank
no. of different words (***types***)
type/token ratios 353
no. of sentences 164 in the text
mean sentence length (in words)
standard deviation of sentence length (in words)
no. of paragraphs 165 in the text
mean paragraph length (in words)
standard deviation of paragraph length (in words)
no. of headings 164 in the text (none here because WordSmith didn't know how to recognise headings)
mean heading length (in words)
no. of sections 165 in the text (here 480 because WordSmith only noticed 1 section per text)
mean section length (in words)
standard deviation of heading length (in words)
numbers 146 removed
stoplist 141 tokens and types removed
the number of 1-letter words
…
the number of n-letter words (to see these scroll the grid horizontally)

(14 is the default 110 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.

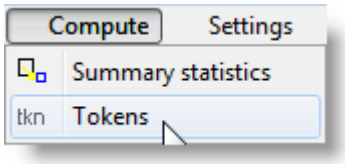
Tokens used for word list

In these data, there were over 2.8 million running words of text, but over 30,000 numbers were not listed separately, so the number of tokens in the word-list is a little under 2.8 million.

MS Word's word count is different!

The number of tokens found is affected by your settings such as treatment of numbers 146, hyphens 544 and mid-word letter settings 146 (e.g. the apostrophe). For that reason you may well find that different programs give different values for the same text. (Besides, in the case of MS Word we are not told how a "word" is defined...)

Sum of entries can be computed after the word-list is created by choosing *Compute* | *Tokens*



and will show the total number of tokens now available by adding the **frequencies** of each entry (you may have deleted some).

See also : WordList display 380 (with a screenshot), Summary Statistics 141, Starts and Ends of Text Segments 164, Recomputing tokens 349.

9.12.1 type/token ratios

If a text is 1,000 words long, it is said to have 1,000 "tokens 533". But a lot of these words will be repeated, and there may be only say 400 different words in the text. "Types 533", therefore, are the different words.

The ratio between types and tokens in this example would be 40%.

Problem with TTR

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.

Standardised TTR

WordList offers a better strategy as well: the *standardised type/token ratio* (STTR) is computed every **n** words as Wordlist goes through each text file. By default 110, 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 370 to any number between 100 and 20,000.

What STTR actually counts

The ratio is computed a) counting every different form 322 as a word (so **say** and **says** are two types) b) using only the words which are not in a stop-list 141 c) those which are within the length you have specified, d) taking your preferences about numbers 554 and hyphens 544 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 524 (1995:131-50) who relies on a notion of 3 stages of accumulation. The WordSmith method of computing STTR was my own invention but parallels one of the methods devised by the mathematician David Malvern working with Brian Richards (University of Reading).

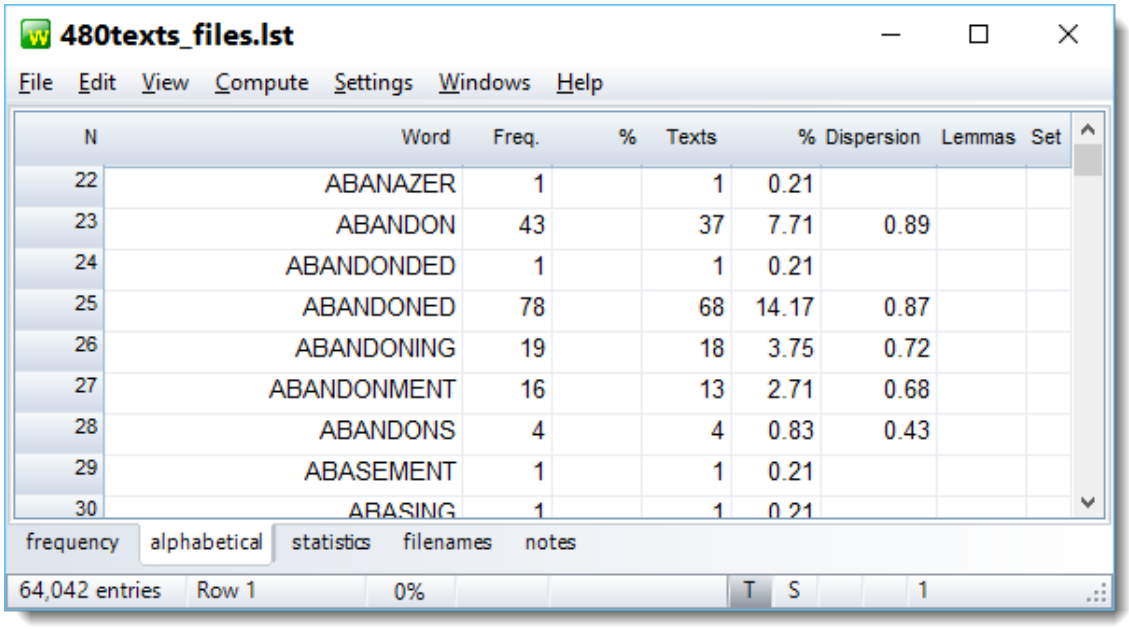
Further discussion

TTR and STTR are both pretty crude measures even if they are often assumed to imply something about "lexical density". Suppose you had a text which spent 1,000 words discussing **ELEPHANT, LION, TIGER** etc, and then 1,000 discussing **MADONNA, ELVIS**, etc., then 1,000 discussing **CLOUD, RAIN, SUNSHINE**. If you set the STTR boundary at 1,000 and happened to get say 48% or so for each section, the statistic in itself would not tell you there was a change involving Africa, Music, Weather. Suppose the boundary between Africa & Music came at word 650 instead of at word 1,000, I guess there'd be little or no difference in the statistic. But what *would* make a difference? A text which discussed clouds and written by a person who distinguished a lot between types of cloud might also use **MIST, FOG, CUMULUS, CUMULO-NIMBUS.** This would be higher in STTR than one written by a child who kept referring to **CLOUD** but used adjectives like **HIGH, LOW, HEAVY, DARK, THIN, VERY THIN** to describe the clouds... and who repeated **DARK, THIN**, etc a lot in describing them.....

(NB. Shakespeare is well known to have used a rather limited vocabulary in terms of measures like these!)

9.12.2 dispersion

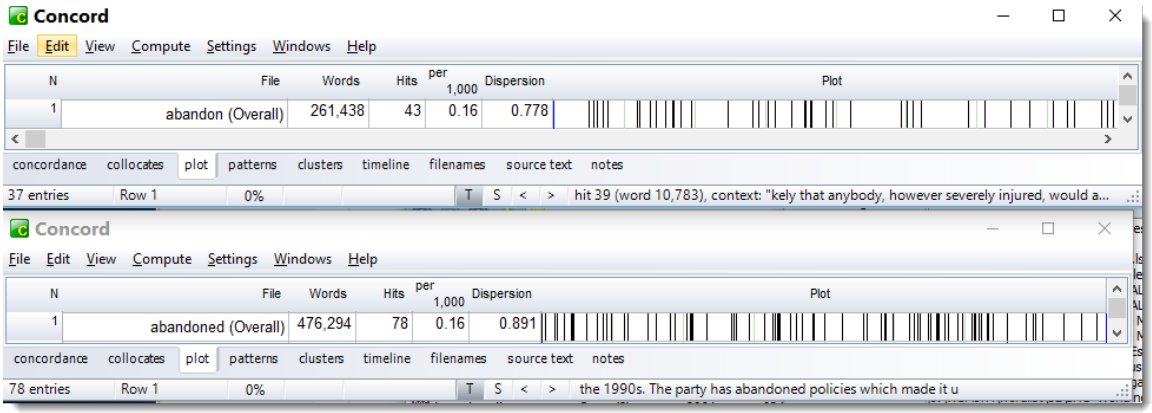
For a word list or for clusters 332 computed from a WordList index, dispersion 554 is calculated 540 using the same formula as for a dispersion plot 230 in Concord. It splits the corpus up into a number of divisions (default = 8) and for every word, computes how the word spreads out in the whole set of texts. It needs to allocate some memory space to the divisions as it works. You can set the number of divisions the procedure uses in the main Controller 367.



The word with highest dispersion in this shot (**ABANDON**) is the one which is most uniformly spread through the corpus. It was present in only 7.71% of the texts. **ABANDONED** came more often and in more texts but was similar in dispersion. A word which is present in nearly all texts such as a grammar word may reach 0.99.

Understanding what the dispersion really shows

The listing above was computed using the whole corpus of 480 texts. If we compare with a dispersion plot in Concord, note



that the dispersion scores are computed differently, using *only the text files in which these individual items appeared* (37 **ABANDON** and 68 for **ABANDONED** -- contrast also the number of *Words* in the two plots). As a result the dispersion values are different from those in a word list over many more texts.

See also: Concord plot dispersion 230

9.12.3 relative entropy

The point of it...

The idea here is similar to dispersion, and it is measured over a series of frequencies such as those in a lemma.

The example given by Gries (2010) is the lemmas **give** and **sing** in English. It is clear from the table that some forms of **sing** (**sings, sang**, and **sung**) are infrequent, but no forms of **give** are very infrequent. The relative entropy value (0.91) for **give** tells you it is more smoothly spread over the 5 verb forms than **sing** is at 0.62.

					Rel. Entropy
give 441	**gives** 105	**giving** 132	**gave** 175	**given** 376	0.91
sing 38	**sings** 2	**singing** 45	**sang** 3	**sung** 2	0.62

Using this function you can therefore compare lemmas in terms of how generally used or specialised all their lemma variants are.

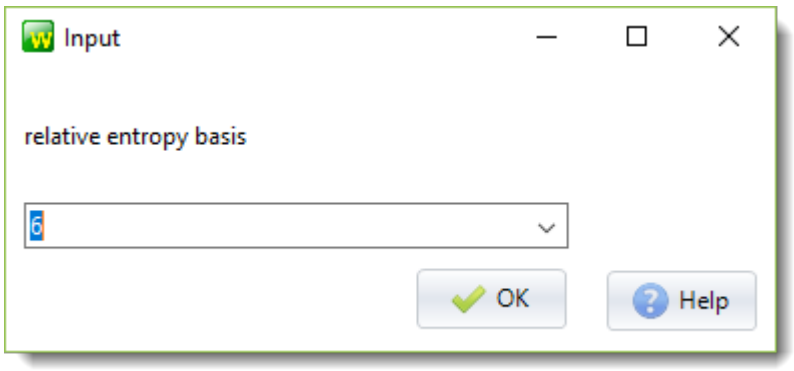
How to compute it

This function suits lemmatised data, so it is not routinely shown for any word list. To compute relative entropy, you will need first to lemmatise 322 at least some word forms in your word list.

Then choose the menu option *Compute* | *Relative entropy*.

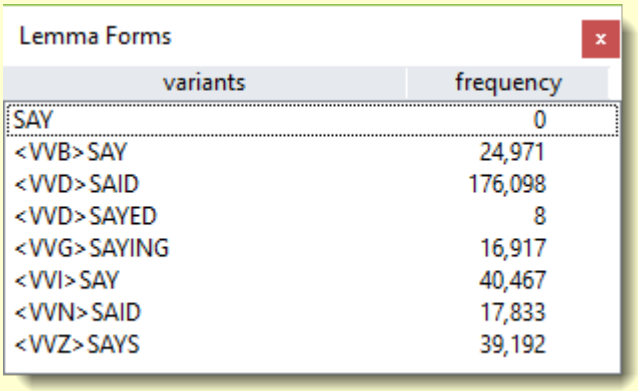
a standard basis for the calculation

Choose a basis for the calculation. Here we choose 6 because there are exactly 6 verb types in the BNC data analysed (not 5 as in Gries's example).



How to choose: comparing like with like

The idea is to compare like with like. There are 6 possible verb forms for each verb, so if a verb has only 4 of the forms we need to add in zero values for two missing forms. Each verb gets a set of 6 values for comparison, therefore.
What if a verb has more than 6? Here is a case.


Lemma Forms

variants	frequency
SAY	0
<VVB>SAY	24,971
<VVD>SAID	176,098
<VVD>SAYED	8
<VVG>SAYING	16,917
<VVI>SAY	40,467
<VVN>SAID	17,833
<VVZ>SAYS	39,192

In the BNC XML version downloaded, it seems that **SAYED** is found and linked to the headword say. (Click here 329 if wondering about the zero frequency for say itself.) To handle this eventuality, all the numbers get sorted and only the top 6 in value get used for our calculation here. Thus eliminating **SAYED**.

So choose a number which will be fitted to each entry. For example if you're computing relative entropy over a set of collocates and if you want to compare results with those of other collocates, choose a maximum number N which will suit most of your collocate lists. Any with more than than N collocates will only use the most frequent N collocates in computing. Any with less with assume some zero frequency collocates.

If you choose zero, all variants of each entry will be used for the calculation.

A new column will get created and compute a relative entropy value for each entry.

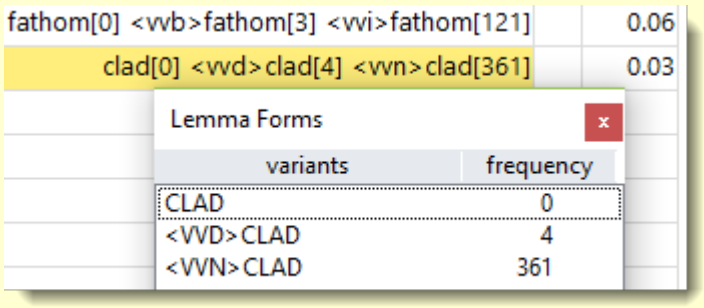
Results

Here is a BNC word list showing only verbs and computed relative entropy. **MAKE** scores higher (0.94) than **SAY** (0.76).

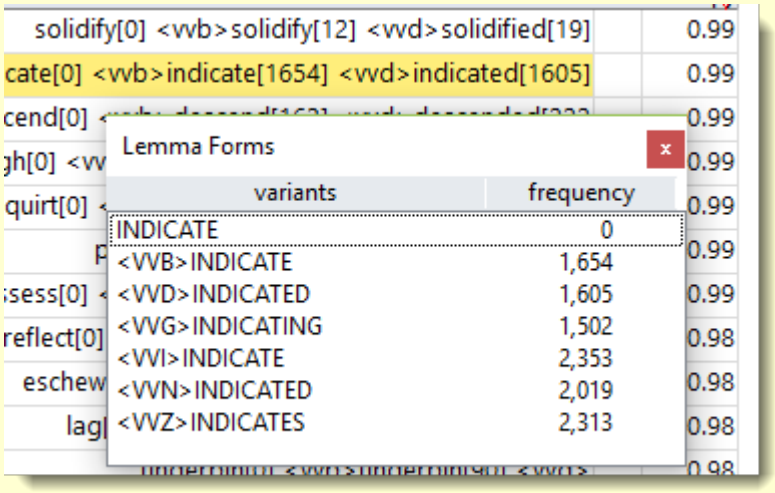
Word	Freq.	%	Texts	%.	Lemmas ...	relative entropy 2
SAY	315,486	0.32	0	.	say[0] <vvb>say[24971]	0.76
GO	224,408	0.23	0	.	go[0] <vvb>go[24986]	0.92
GET	207,797	0.21	0	.	get[0] <vvb>get[28474]	0.90
MAKE	197,327	0.20	0	.	make[0] <vvb>make[18224]	0.94
SEE	184,701	0.19	0	.	see[0] <vvb>see[56127]	0.84
KNOW	176,796	0.18	0	.	know[0] <vvb>know[67990]	0.83
~~<VVD>SAID~~	~~176,098~~	~~0.18~~	~~3,302~~	~~3...5.~~		
TAKE	172,139	0.17	0	.	take[0] <vvb>take[18391]	0.94
COME	143,031	0.15	0	.	come[0] <vvb>come[23536]	0.95
THINK	140,934	0.14	0	.	think[0] <vvb>think[53982]	0.82

High and low Relative Entropy scores

Moving to the *Alphabetical tab* in order to sort on the *Relative Entropy* column, we can see what sorts of verbs score highly and which much lower: CLAD and FATHOM are at the bottom of the list. As you see, CLAD has only two of the 6 forms, the past-participle being much more frequent than the simple past.



At the top we get **SOLIDIFY, INDICATE, DESCEND**. Whether high or low frequency verbs overall, these all have much more uniformly spread frequencies.

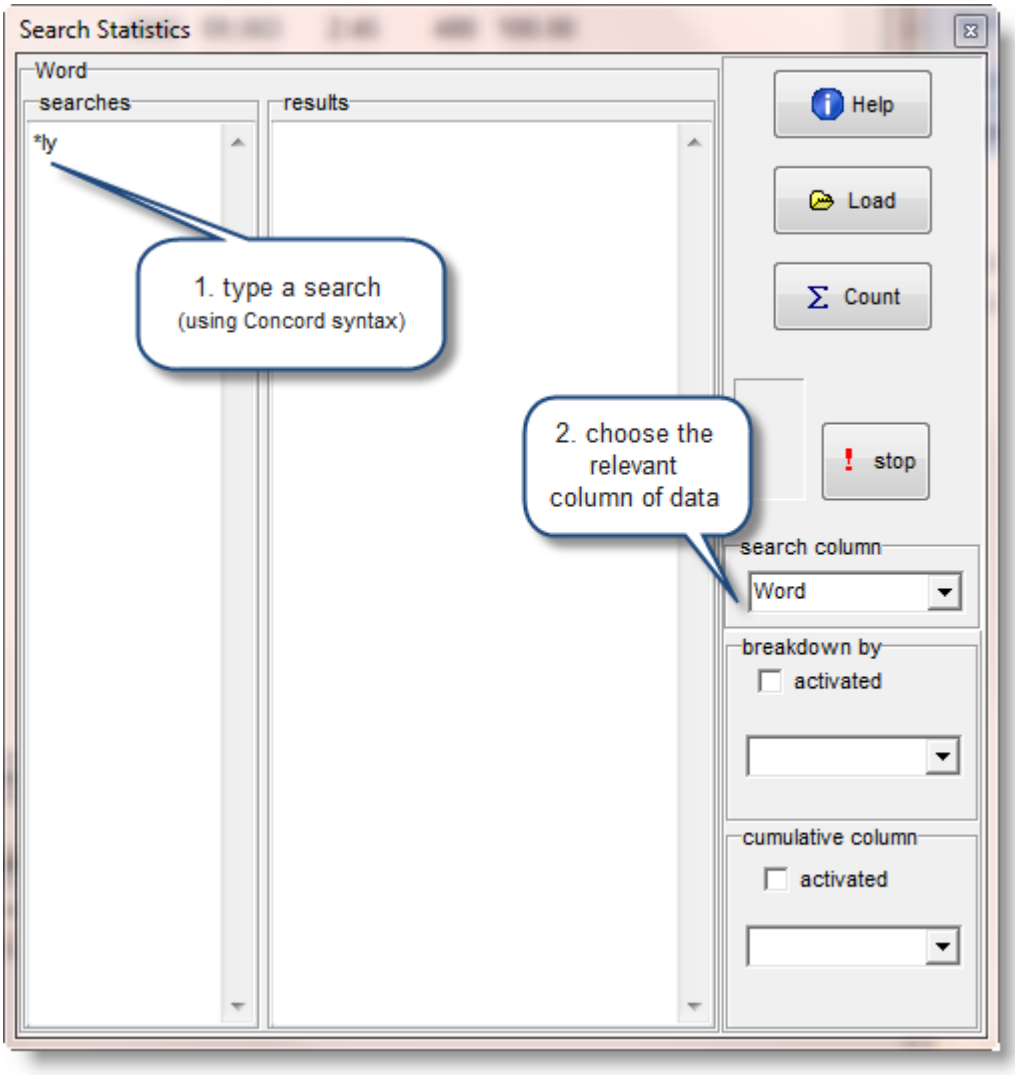


See also: collocates and relative entropy 209, relative entropy formula 540

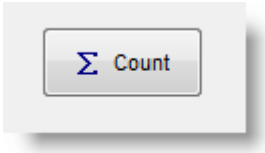
9.12.4 summary statistics, breakdowns

A word list's statistics give you data about the corpus, but you may need more specific information about individual words in a word list too.

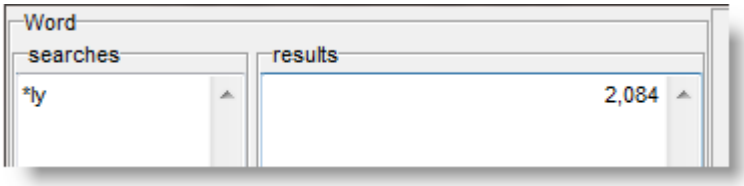
How many end in -ly?



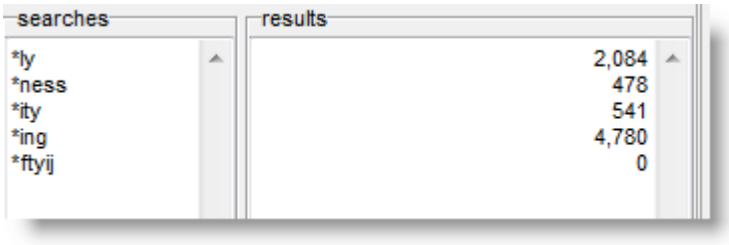
Press Count



to get something like this:

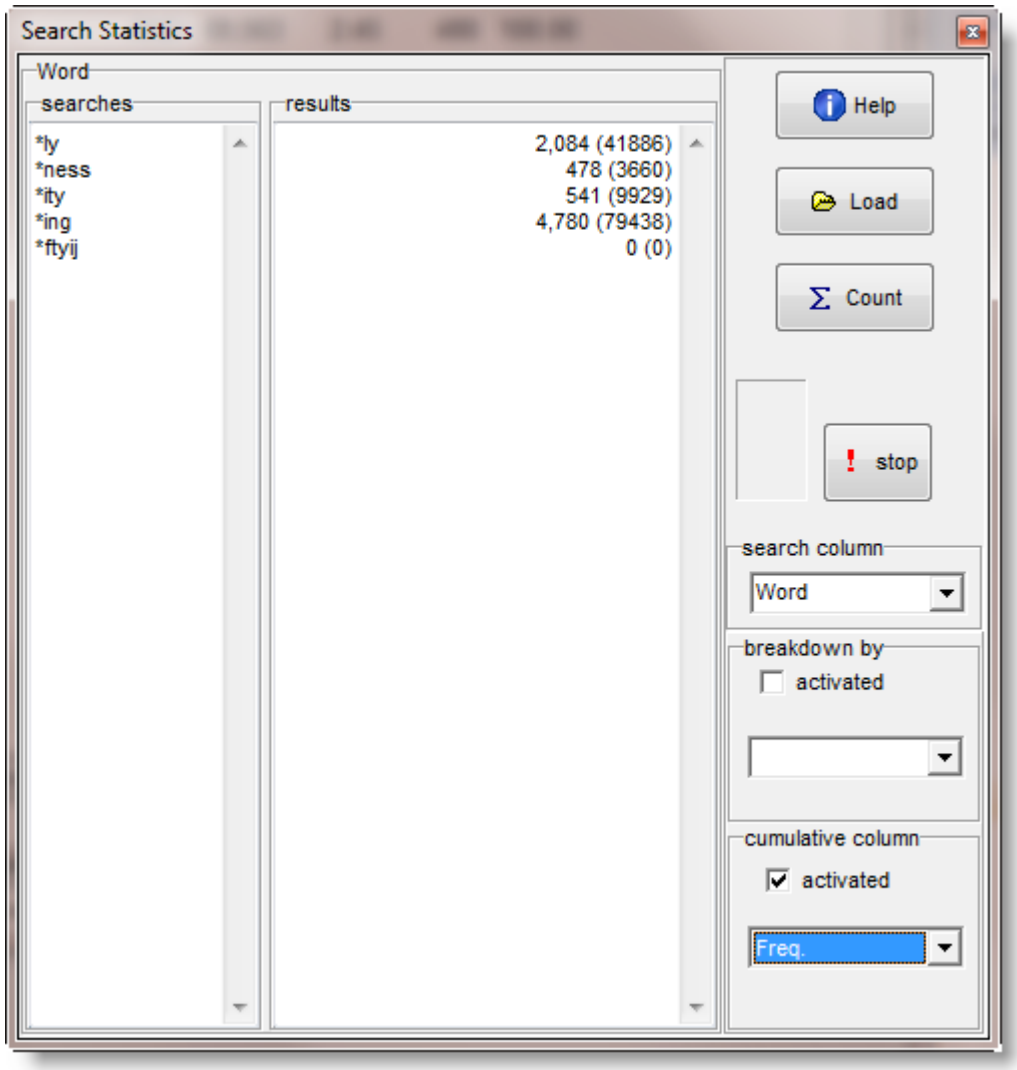


There is no limit on the searches:



Cumulative Column

A cumulative count adds up scores on another column of data apart from the one you are processing for your search. The columns in this window are for numerical data only. Select one and ensure *activated* is ticked.



In this example, a word-list was computed and a search was made of 4 common word endings (and one ridiculous one). For **-LY** there are 2,084 types, with a total of 41,886 tokens in this corpus. **-ITY** and **-NESS** are found at the ends of fairly similar numbers of word-types, but **-ITY** has many more tokens in these data.

Breakdown

See the example for Concord 138

Load button

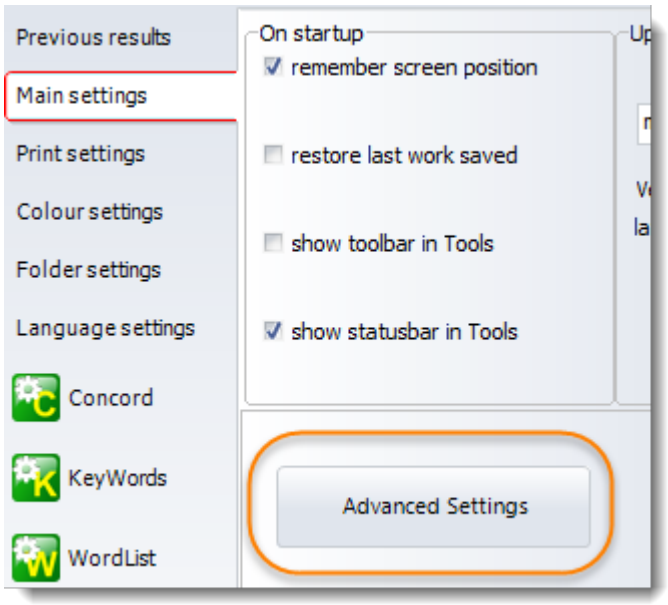
see the explanation for count data frequencies 141.

9.13 stop-lists and match-lists

In WordList, a stop list 141 is used in order to filter out some words, usually high-frequency words, that you want excluded from your word-list. The idea of a match-list 82 is to be able to compare all the words in your word list with another list in a plain text file and then do one of a variety of operations such as deleting the words which match, deleting those which don't, or just marking the ones in the list.

For both, you can define your own lists and save them in plain text files.

Settings are accessed via the WordList menu or by an Advanced Settings button in the Controller



Stop list and clusters

If you compute clusters using an Index, any WordList stop list words can be excluded. As you start to compute clusters you will be asked whether to remove stop list words within clusters. So if **THE** is in your stop list and you answer *Yes*, a cluster like **BEAT THE RAP** would get cut out.

See also: lemma lists 322, general explanation of stop-lists 141

9.14 import words from text list

the point of it

You might want a word list based on some data you have obtained in the form of a list, but whose original texts you do not have access to.

requirements

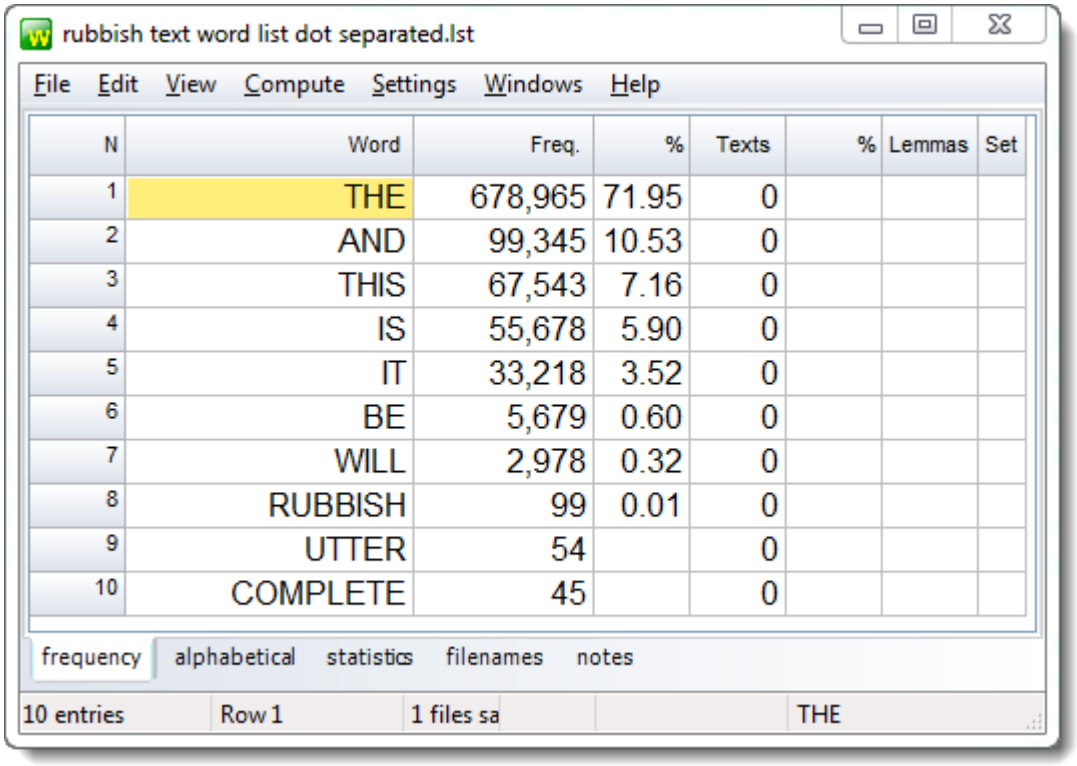
Your text file can be in any language 77 (select this before you make the list), and can be in Unicode or ASCII or ANSI, plain text.
<Tab> characters are expected to separate the columns of data. Decimal points and commas will be ignored. Words will have leading or trailing spaces trimmed off. The words do not need to be in frequency or alphabetical order. You need at least a column with words and another with a number representing each word's frequency.

example

```
; My word list for test purposes.

THIS      67,543
IT        33,218
WILL      2,978
BE        5,679
COMPLETE  45
AND       99,345
UTTER     54
RUBBISH   99
IS        55,678
THE       678,965
```

You should get results like these.

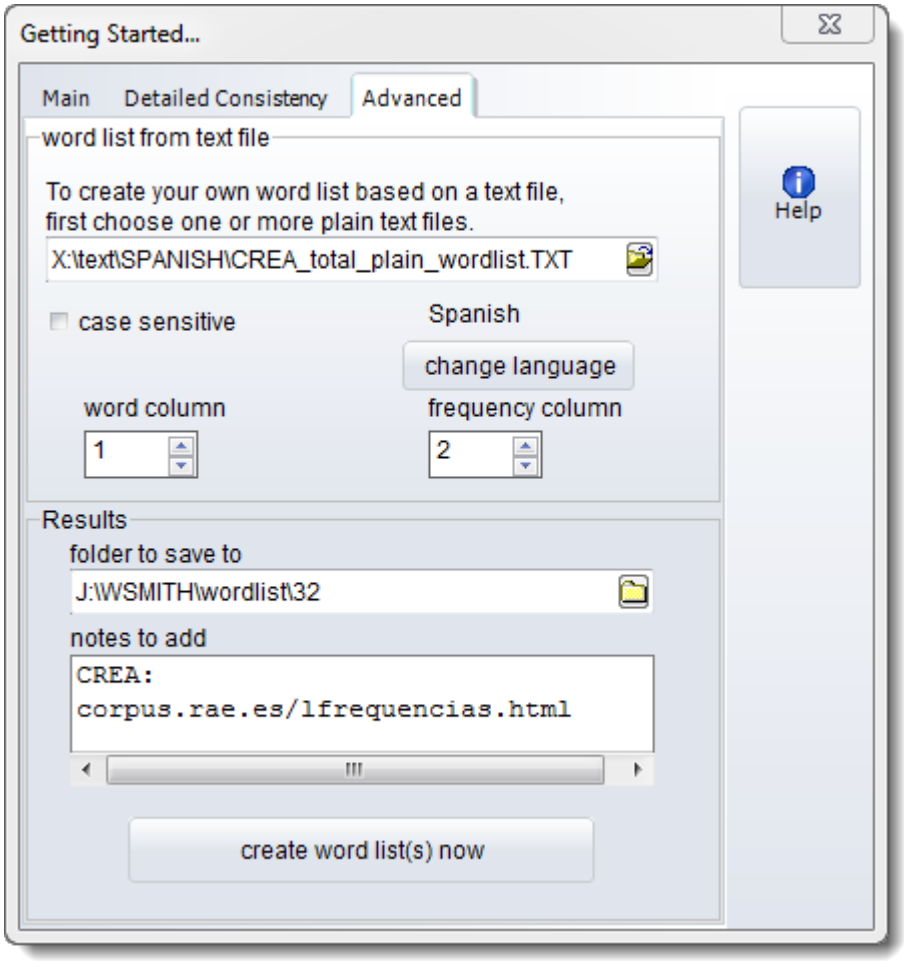


Statistics are calculated in the simplest possible way: the word-lengths (plus mean and standard deviation), and the number of types and tokens. Most procedures need to know the total number of running words (tokens) and the number of different word types so you should manage to use the word-list in KeyWords etc.

The total is computed by adding the frequencies of each word-type (`67543+33218+2978` etc. in the example above).
Optionally, a line can start **`\TOTAL=\`** and contain a numerical total, eg.
`\TOTAL=\  299981`
In this case the total number of tokens will be assumed to be 299981, instead.

how to do it

When you choose the *New* menu option ( ) in WordList you get a window offering three tabs: a *Main* tab for most usual purposes,



one for *Detailed Consistency* 311, and another (*Advanced*) for creating a word list using a plain text file.

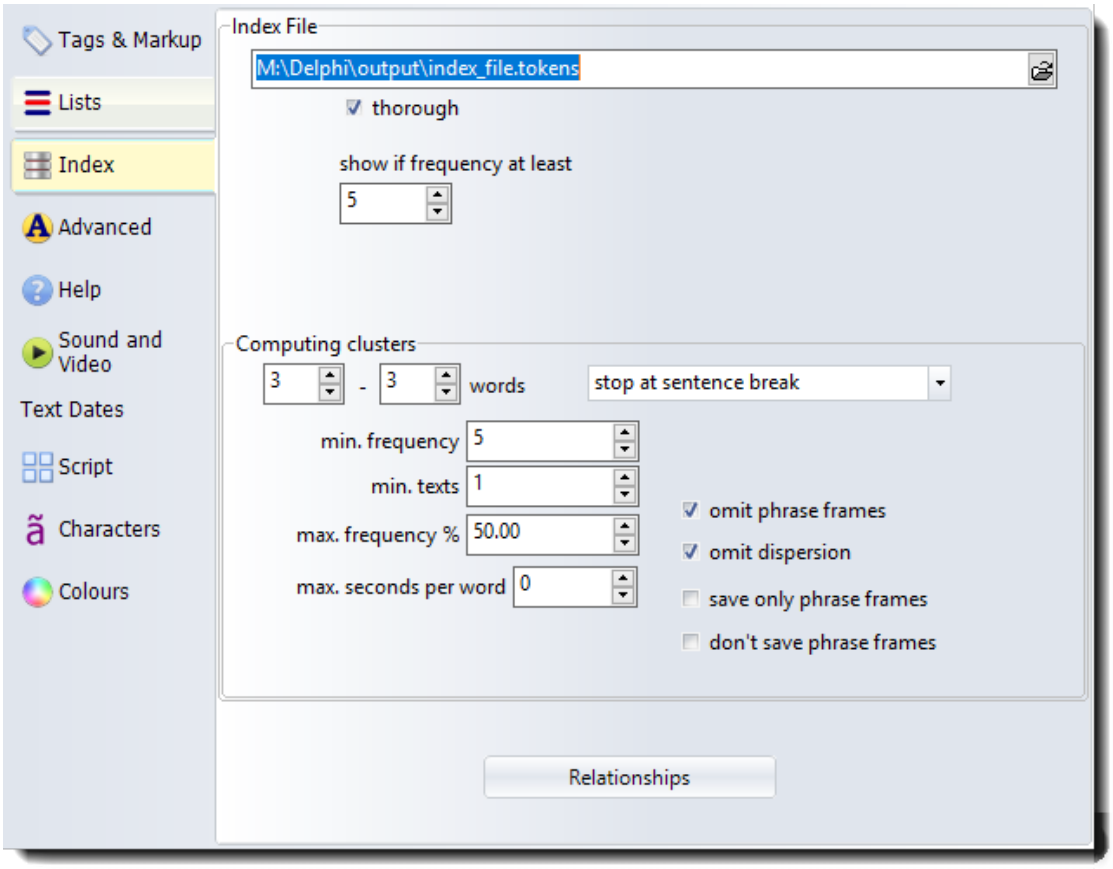
Set the *word column* and *frequency column* appropriately according to the tabs in each line. (Column 1 assumes that the word comes first before any tabs; in the case of CREA's Spanish word-list there is a column for ranking so the word and frequency columns would need to be 2 and 3 respectively.)

Choose your .txt file(s) and a suitable folder to save to, add any notes you wish, and press *create word list(s) now*.

9.15 settings

There are two places where settings for functions for the WordList tool can be found, in the main Controller 367 and in the Index settings page 366.

9.15.1 WordSmith controller: Index settings



Index File

The filename is for a default index which you wish to consider the usual one to open.

thorough concordancing: when you compute a concordance from an index, you will either get (*thorough* checked) or not get (if not checked) full sentence, paragraph and other statistics 198 as in a normal concordance search. (Computing these statistics takes a little longer.)

show if frequency at least: determines which items you will see when you load up the index file. (What you see looks like a word list but it is reading the underlying index.)

Clusters

the minimum and maximum sizes are 2 and 8. Set these before you compute a multi-word word list 332 based on the index. A good maximum is probably 4 or 5.

min. frequency: minimum frequency for each cluster.

min. texts: minimum number of texts a cluster must be in.

stop at: you can choose where you want cluster breaks to be assumed. With the setting above (no limits), "**`I wrote the letter. Then I posted it`**" would consider **`letter then I posted`** as a possible multi-word string even though there's a sentence break 217 between them.

max. frequency %: saves processing very high frequency words. (**THE** in most English text takes up about 5%, so a setting of 3% will save computing clusters which start with the topmost frequency words.)
max. seconds per word: a setting allowing you to ignore high-frequency items which occur so often that procesing slows down. (0=no limit)
omit phrase frames: (default = checked) avoids showing phrase frames 336 in a word cluster list.
omit dispersion: (default = not checked) avoids computing dispersion scores 354 in an index.
save only phrase frames: (applies when saving a word list of clusters)
don't save phrase frames: (applies when saving a word list of clusters)

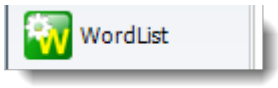
Relationships

See relationships computing 344.

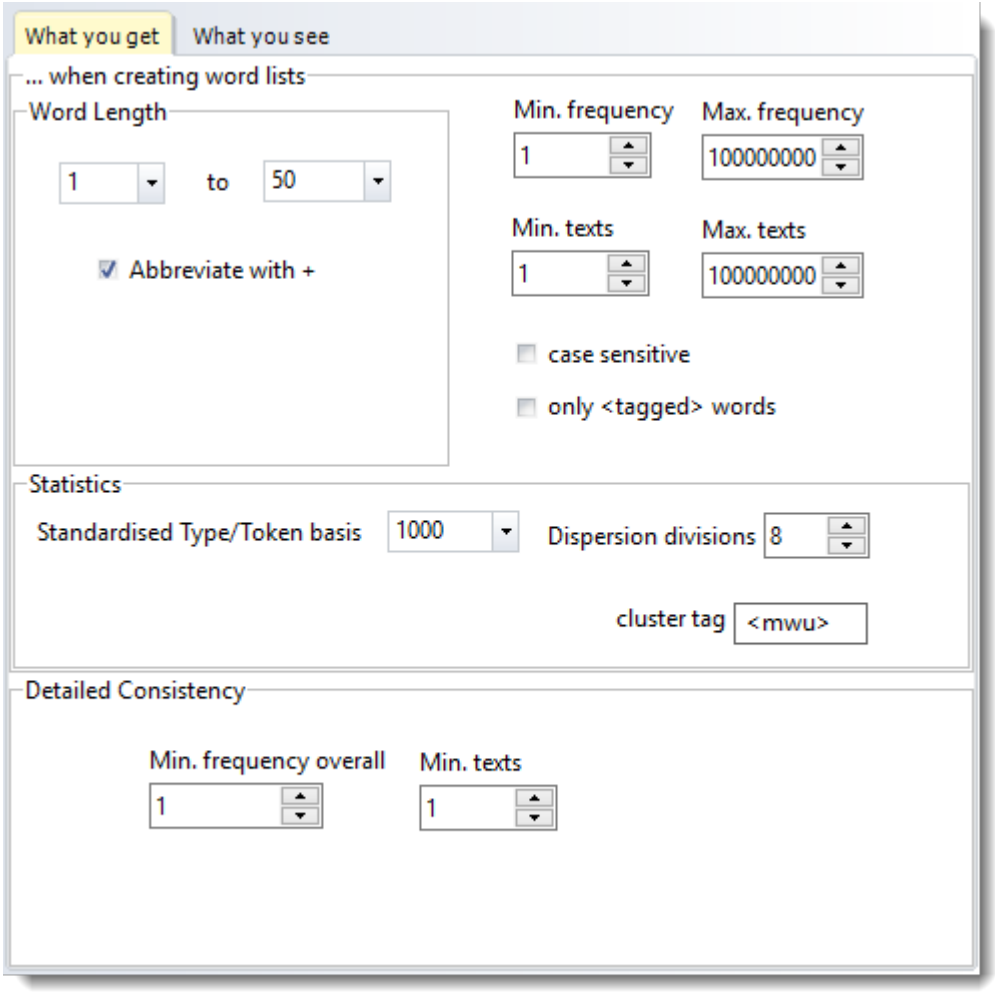
See also: WordList index 330

9.15.2 WordSmith Controller WordList settings

9.15.2.1 what you GET



These are found in the main Controller 5 marked *WordList* .
This is because some of the choices -- e.g. Minimum & Maximum Settings 370 -- may affect other Tools.



Word Length & Frequencies

See Minimum & Maximum Settings 370.

Case Sensitive

Normally, you'll make a case-insensitive word list. If you wish to make a word list which distinguishes between **the**, **The** and **THE**, turn on case sensitivity 371 here. *Note* : Stop 141, match 82 and lemma 444 lists can also be case sensitive.

Only <tagged> words: Omits all words which don't have a partner tag 187.

Standardised Type/Token

See WordList Type/Token Information 353.

Dispersion divisions

This setting applies to the Dispersion 354 column in a word list. By default 8 but you can set the number to 0 to speed up processing. Avoid a number higher than 25 as this will cause memory strain.

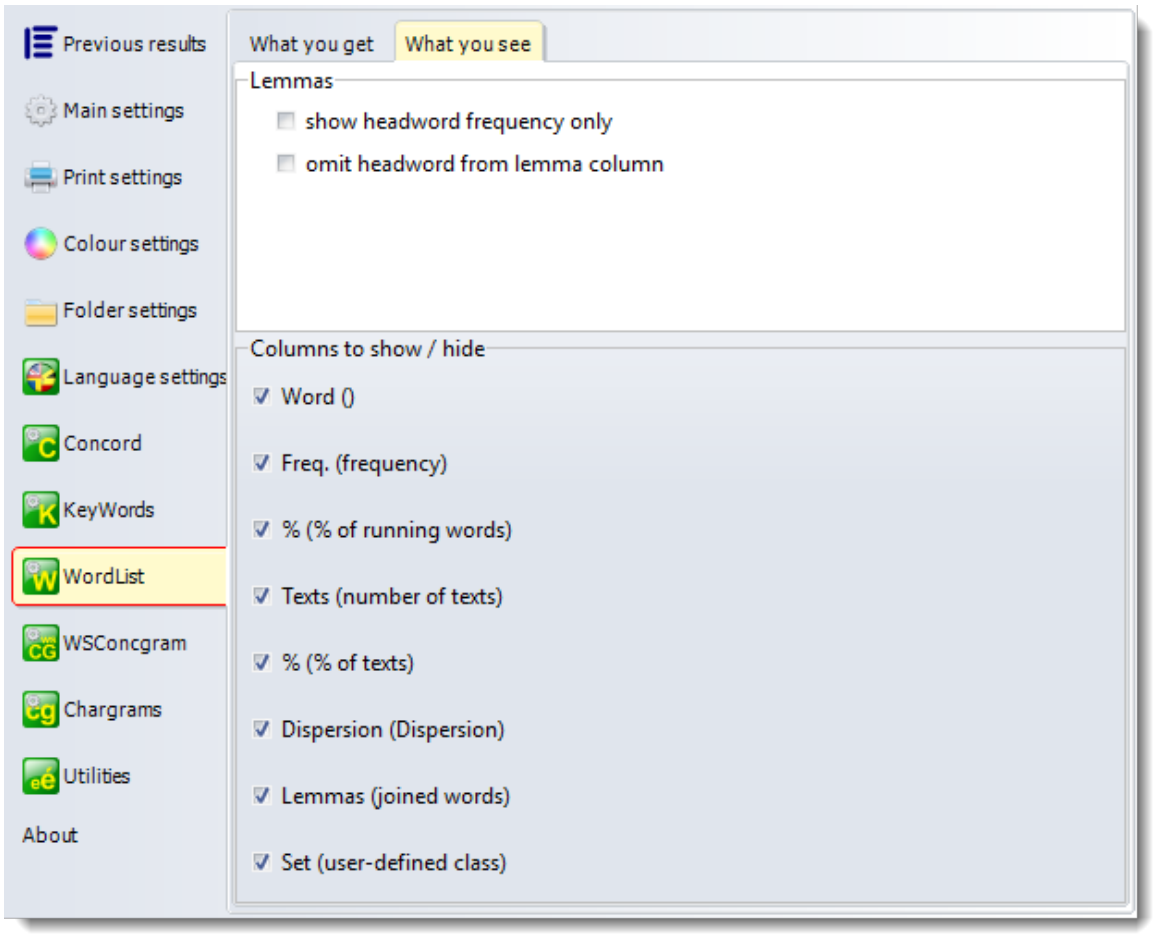
Cluster Tag: used to include some multi-word items in a word list 306.

Detailed Consistency

Min. frequency overall = a total frequency for a word to be included in the Detailed Consistency 311 list.
Min. texts = the minimum number of texts that word must appear in.

See also: Using Index Lists 330, Viewing Index Lists 338, WordList Help Contents 305, WordList and tags 374, Computing word list clusters 332.

9.15.2.2 what you SEE



Lemma Visibility

By default in a word-list you'll see the frequency of the headword plus the associated forms; if you check the *show headword frequency only* box, the frequency column will ignore the associated

wordform frequencies. Similarly, if you check *omit headword from lemma column* you will see only the associated forms there.

Columns

The *Columns to show/hide* list offers all the standard columns: you may uncheck ones you normally do not wish to see. This will only affect newly computed data: earlier data uses the column visibility, size, colours etc. already saved. They can be altered using the Layout 60 menu option at any time.

9.15.3 minimum & maximum settings

These include:

minimum word length

Default: 1 letter. When making a word-list, you can specify a minimum word length, e.g. so as to cut out all words of less than 3 letters.

maximum word length

Default: 49 letters. You can allow for words of up to 50 characters in length. If a word exceeds the limit and Abbreviate with + is checked, WordList will append a + symbol at the end of it to show that it was cut short. (If Abbreviate with + is not checked, the long word will be omitted from your word list. You might wish to use this to set both minimum and maximum to say, 4, and leave Abbreviate with + un-checked – that way you'll get a word-list with only the 4-letter words in it.

minimum frequency

Default: 1. By default, all words will be stored, even those which occur once only. If you want only the more frequent words, set this to any number up to 32,000.

maximum frequency

Default maximum is 100 million. You might set this to say 500, and the minimum to 50: that way your word-list would hold only the moderately common words.

min. texts

This setting lets you specify a minimum number of texts each word must appear in (default = 1 text).

max. texts

This setting lets you specify the maximum number of texts each word may appear in (default = 32,000).

only words with defined tags 187

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 353.

See also: Text Characteristics 145, Stop Lists 141, Setting Defaults 110

9.15.4 case sensitivity

Normally, you'll make a case-insensitive word list, especially as in most languages capital letters are used not only to distinguish proper nouns but also to signal beginnings of sentences, headings, etc. If, however, you wish to make a word list which distinguishes between major, Major and MAJOR, turn on case sensitivity 367 (in the Controller 5). That way you will get a word list formatted as the words appeared in the texts:

N	Word	Freq.	%	Texts	% Disp...ion	Lemmas	Set
1	the	129,206	5.51	322	99.69	0.84	
2	of	65,910	2.81	322	99.69	0.84	
3	to	63,359	2.70	322	99.69	0.84	
4	and	54,195	2.31	322	99.69	0.85	
5	a	48,894	2.09	322	99.69	0.85	
6	in	41,394	1.77	322	99.69	0.84	
7	that	31,542	1.35	321	99.38	0.76	
8	is	27,076	1.15	322	99.69	0.83	
9	#	26,061	1.11	272	84.21	0.69	
10	I	23,146	0.99	296	91.64	0.64	
11	it	20,472	0.87	321	99.38	0.74	
12	for	20,236	0.86	322	99.69	0.84	
13	you	18,399	0.78	288	89.16	0.55	
14	on	17,530	0.75	322	99.69	0.84	
15	was	16,105	0.69	321	99.38	0.85	
16	be	15,474	0.66	321	99.38	0.82	
17	with	13,988	0.60	322	99.69	0.85	
18	as	13,453	0.57	322	99.69	0.84	
19	have	12,798	0.55	321	99.38	0.80	
20	The	12,792	0.55	321	99.38	0.78	
21	at	11,898	0.51	322	99.69	0.84	
22	are	11,347	0.48	321	99.38	0.83	
23	by	10,736	0.46	322	99.69	0.80	
24	not	10,185	0.43	320	99.07	0.83	
25	we	9,970	0.43	297	91.95	0.61	
26	er	9,921	0.42	97	30.03	0.45	

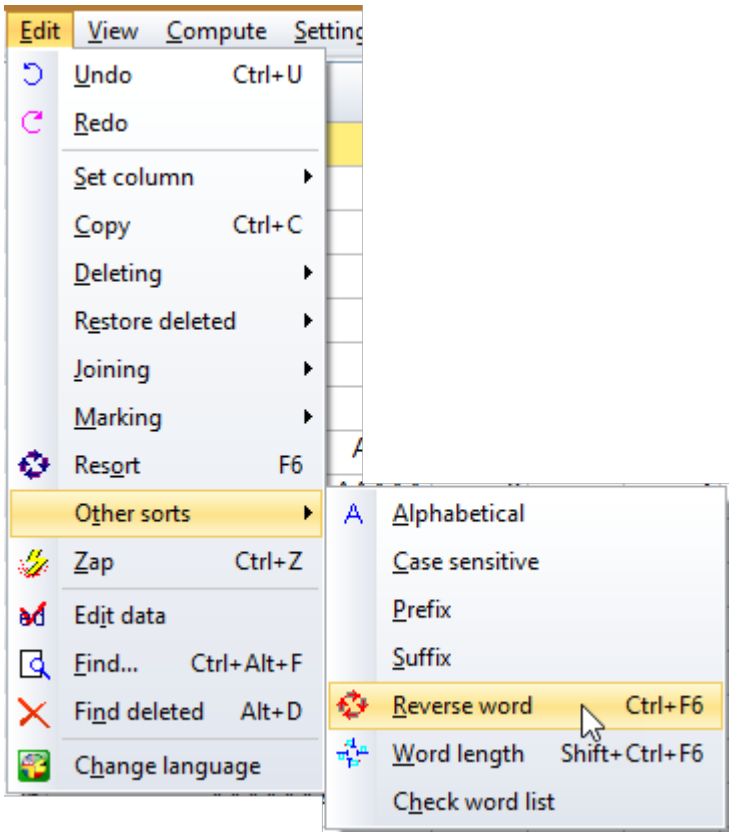
Press Ctrl+L or choose the *Layout* 60 menu option ( ) to change case in how you see most lists.

9.16 sorting

How to do it...

Sorting can be done simply by pressing the top row of any list. Press again to toggle between ascending & descending sorts.

With a word-list on your screen, the main Frequency window doesn't sort, but you can re-sort the Alphabetical window (look at the tabs at the bottom of WordList to choose the tab) in a number of different ways.



Resorting a word list

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 77 for the words being sorted. Spanish, for example, uses this order:
A,B,C,CH,D,E,F,G,H,I,J,K,L,LL,M,N,Ñ,O,P,Q,R,S,T,U,V,W,X,Y,Z.

KeyWords and other comparisons require an alphabetically-ordered list in ascending order. If you get problems, please open the word lists in WordList, choose the "alphabetical" tab, sort by pressing the "Word" header until the sort is definitely alphabetical ascending, then choose the Save menu option.

Reverse Word Sort (Ctrl+F6)

This is so that you can sort words by suffix. The order is determined by word endings, not word beginnings. You will therefore find all the **`-ing`** forms together.

Word Length Sort (Shift+Ctrl+F6)

This is so that you can sort words by their length (1-letter, 2-letter, etc up to 50-letter words) Within a set of equal-length words, there's a second, alphabetical sort.

Consistency Sort

Press the "Texts" header to re-sort the words according to their consistency 311.

Check word list

Occasionally a word list may be found to contain items like these

51,113	DÆMONS	3
51,114	DAEMONS	9

with words which share the same alphabetical order but with slightly different spelling. This may also be because of some formatting problem, as here:

213,497	ÞAN	6
213,498	THAN	280,572

where the word THAN and ÞAN likewise share the same alphabetical rank order (even though ÞAN is certainly a typo or other error in this case).

A word list should ideally not contain words of equal rank as that makes computing key words trickier (the key words procedure could think that the word THAN has a frequency of only 6). To check whether you have this issue, choose the *Check word list* option in the menu.

If you do, and get a list shown of such pairs like this

variants

Word 1	Word 2
ASBAEK	ASBÆK
DÆMONS	DAEMONS
HÆMORRHAGING	HAEMORRHAGING
KJÆRGAARD	KJAERGAARD
ORTHOPÆDICS	ORTHOPAEDICS
TH	Þ
ÞA	THA
ÞAN	THAN
ÞAND	THAND

See also: Concord sort 238, KeyWords sort 296, Editing entries 75; Accented characters 527; Choosing Language 77

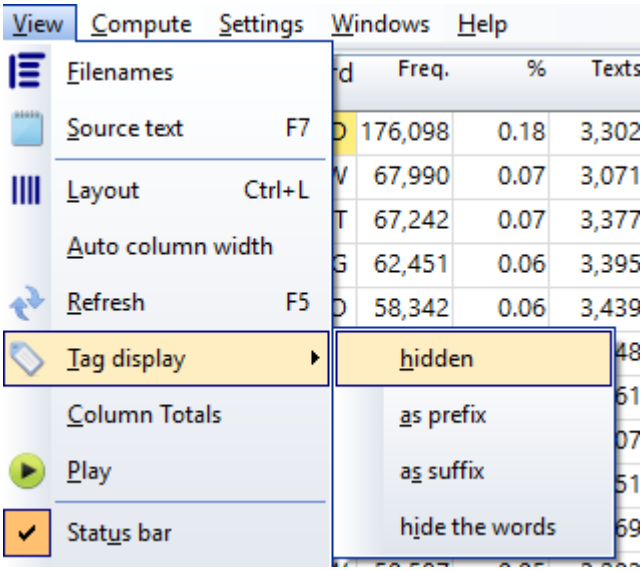
9.17 WordList and tags

If you have defined a tag file and made the appropriate settings 367 to load it, you can get a word-list which displays tags and words together as in this example, where the tag is viewed as if it were a prefix.

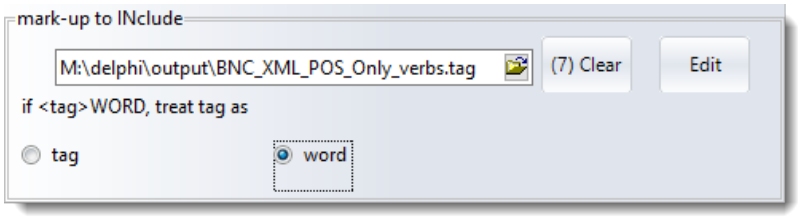
This function operates where your corpus has tags immediately to the left of words, as in **`...`** **`<noun>table`** **`...`** for example.

A word list with neighbouring tags as if they were prefixes?

Choose whether you want only the tags, only the words or both in *View* | *Tag display*:

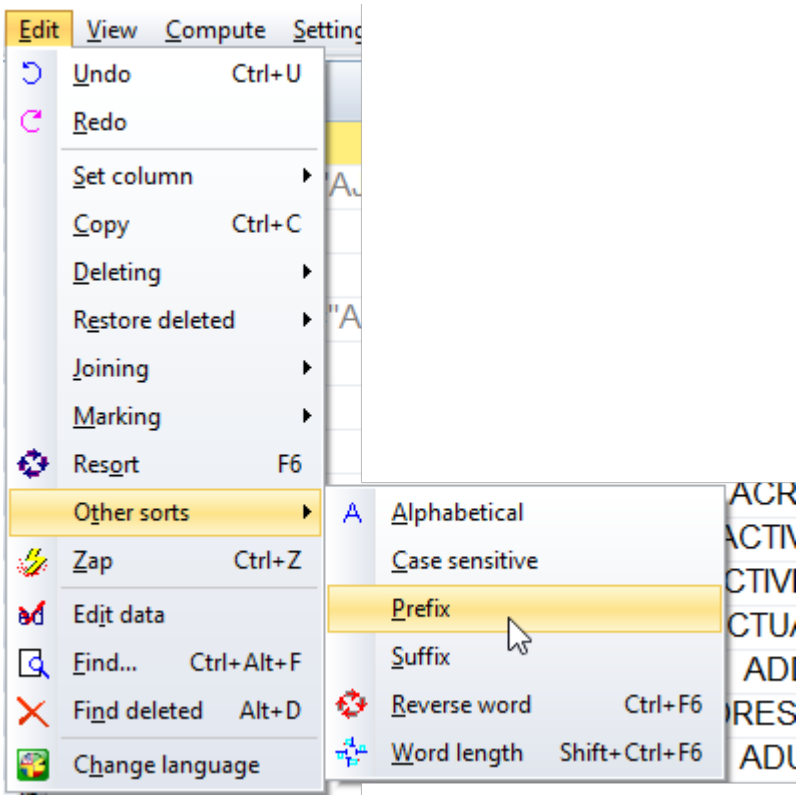


hidden	shows only the word
as prefix	shows the tag as if it were a prefix
as suffix	puts the tag at the other end of the word
hide the words	treats the tag as the word wanted, useful if you want a part-of-speech word list without the words, for example. This function requires word lists made after loading up a tag file 172 with such tags set to be treated as words.



N	Word	Freq.	%
406	<w NN2>BEDROOMS	1	
407	<w NN2>BEDS	1	
408	<w VBN>BEEN	30	0.09
409	<w NN1>BEESWAX	1	
410	<w PRP>BEFORE	27	0.08
411	<w CJS>BEFORE	9	0.03
412	<w VVB>BEGIN	2	
413	<w VVG>BEGINNING	1	
414	<w AV0>BEHIND	1	
415	<w PRP>BEHIND	6	0.02
416	<w AJ0>BEIGE	1	
417	<w VBG>BEING	29	0.09
418	<w NN1>BELL	5	0.01
419	<w AV0>BELOW	10	0.03
420	<w PRP>BELOW	4	0.01
421	<w NN1>BENCH	14	0.04
422	<w VVI>BEND	1	

In its *Alphabetical* view, the list can be sorted on the tag or the word.



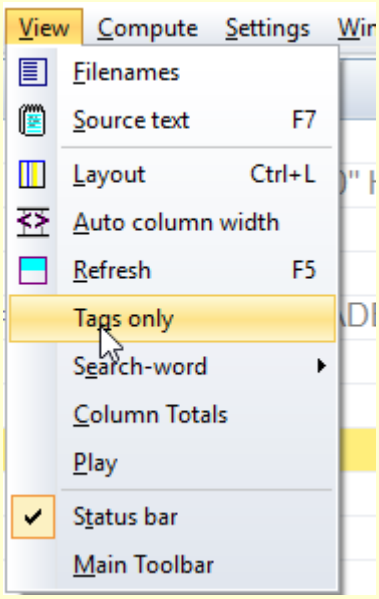
Tag filtering

The point of it

You might want to filter your word list for example removing all entries except adjectives. This word list based on BNC XML text has 98 words more frequent than the top frequency adjective, **young**. The objective was to get rid of all the words except the ones with prefixes (here only adjectives but there could be several different parts of speech tagged).

N	Word	Freq.
95	MAN	10
96	ME	10
97	MY	10
98	WIFE	10
99	<W C5="AJ0" HW="YOUNG" POS="ADJ">YOUNG	10
100	DIE	9
101	FOUR	9
102	GO	9
103	HIMSELF	9

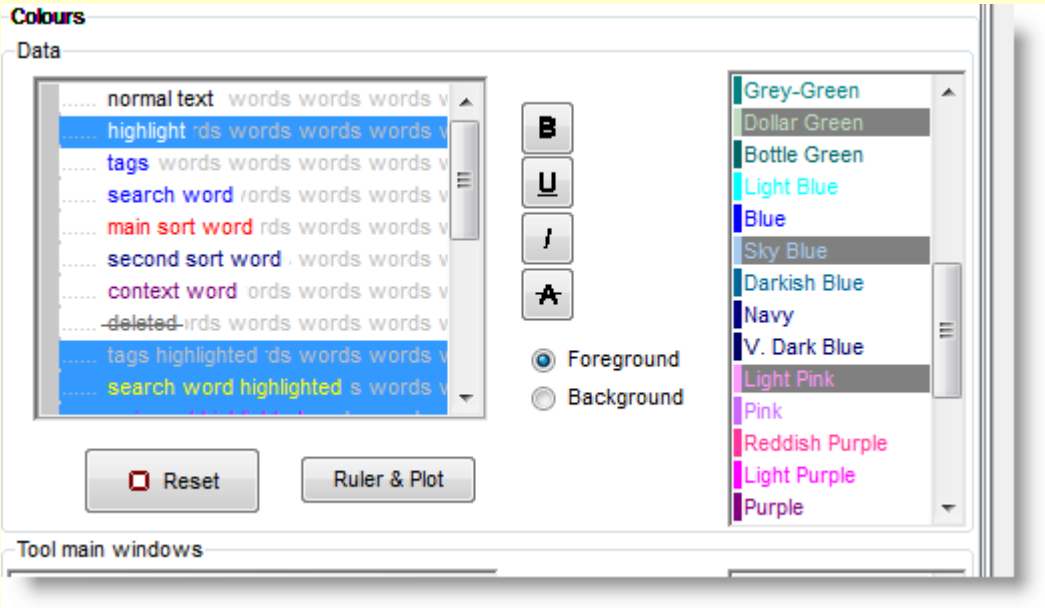
How to do it



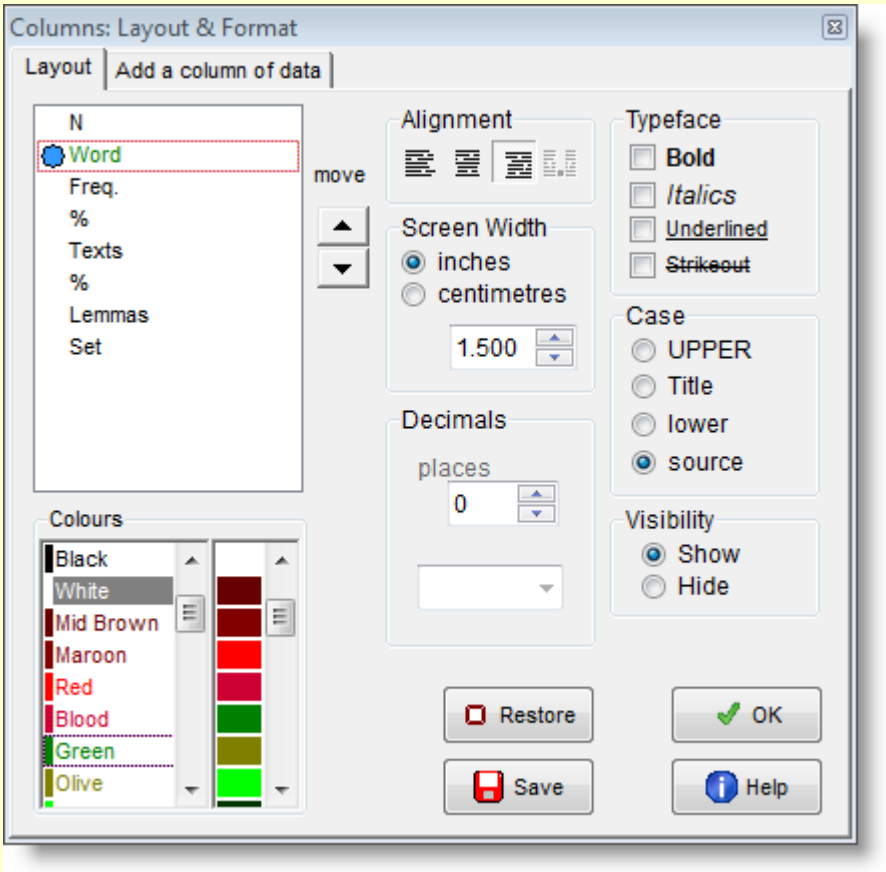
The tags only option will alternately mark all words with tag prefixes as deleted (you can later zap them) or mark all except the tag-prefix words as deleted.

N	Word	Freq.	%
~~136~~	~~BERKSHIRE~~	~~1~~	~~0.02~~
~~137~~	~~BESIDE~~	~~1~~	~~0.02~~
138	<W C5="AJS" HW="GOOD" POS="ADJ">BEST	2	0.03
~~139~~	~~BETTER~~	~~1~~	~~0.02~~
140	<W C5="AJC" HW="GOOD" POS="ADJ">BETTER	1	0.02
~~141~~	~~BETWEEN~~	~~3~~	~~0.05~~
~~142~~	~~BIBLE~~	~~8~~	~~0.14~~
143	<W C5="AJ0" HW="BIBLICAL" POS="ADJ">BIBLICAL	1	0.02
144	<W C5="AJ0" HW="BIG" POS="ADJ">BIG	2	0.03
145	<W C5="AJS" HW="BIG" POS="ADJ">BIGGEST	1	0.02
~~146~~	~~BIT~~	~~2~~	~~0.03~~
147	<W C5="AJ0" HW="BIZARRE" POS="ADJ">BIZARRE	1	0.02
148	<W C5="AJ0" HW="BLACK" POS="ADJ">BLACK	2	0.03

How to colour them



Then in WordList, I chose *View* | *Layout* as in this screenshot, selected the Word column header and chose green below.



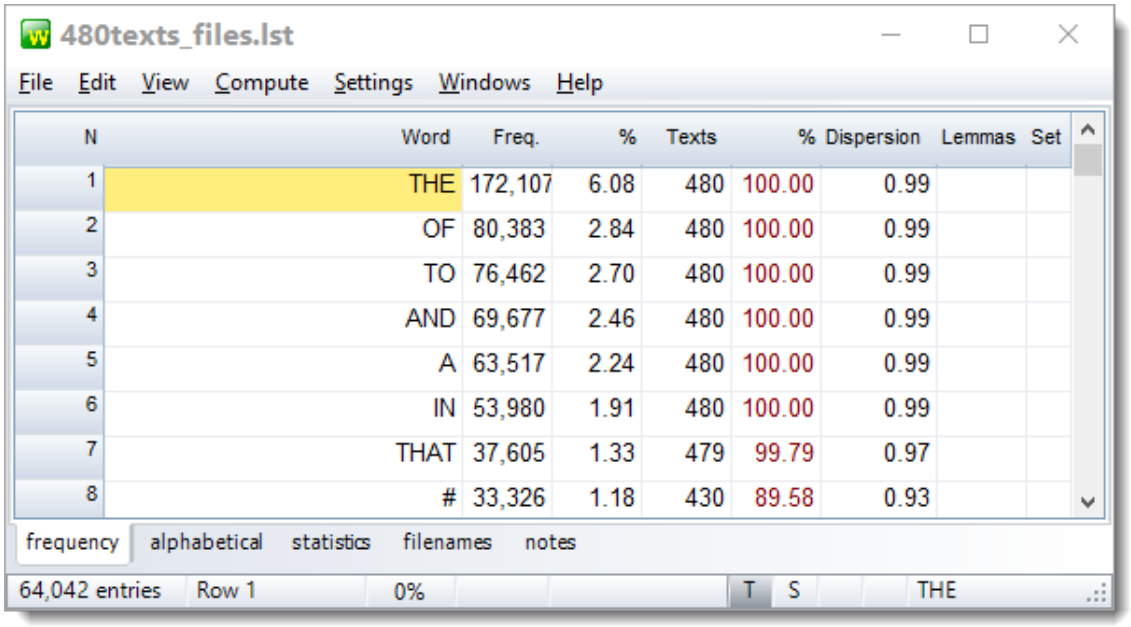
To colour these as in the example, in the main Controller I chose Blue for the foreground for tags (as the default is a light grey).

9.18 WordList display

Each WordList display shows

- the word
- its frequency
- its frequency as a percent of the running words in the text(s) the word list was made from
- the number of texts each word appeared in
- that number as a percentage of the whole corpus of texts
- the word's dispersion 354

The Frequency display might look like this:



Here you see the top 8 word-types in a word list based on 480 texts. There are 172,107 occurrences of these words (tokens 533) altogether. The Freq. column shows how often each word cropped up (**OF** appeared 80,383 times in the 480 texts), and the % column tells us that the frequency represents 2.84% of the running words in those texts. The Texts column shows that **OF** comes in 480 texts, that is 100% of the texts used for the word list. It has a dispersion value of .99 showing that it is a word spread very evenly through the whole corpus of 480 texts.

Per 100, 1000 or million?

In advanced settings you can display the % column 131 in different formats:

gs Windows Help

Word	Freq.	per 1000 words	Texts
AND	664	27.55	1
THE	658	27.30	1

Another thing to note is that there seems to be a word #, with over 33 thousand occurrences. That represents a number or any word with a number in it such as **EX658**. If you want to see all such forms, change the number setting 146.

480texts_files.lst

File Edit View Compute Settings Windows Help

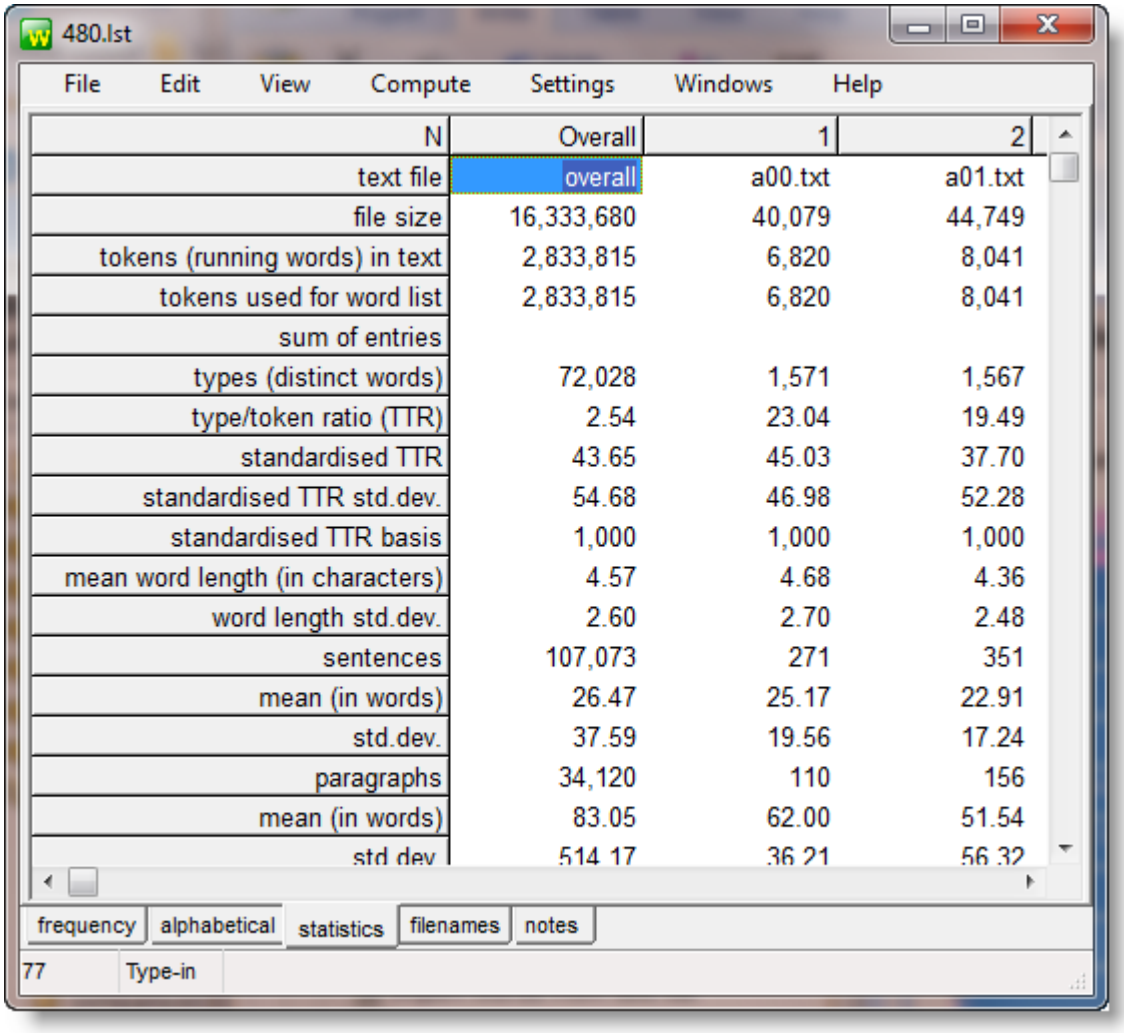
N	Word	Freq.	%	Texts	%	Dispersion	Lemmas	Set
22	ABANAZER	1		1	0.21			
23	ABANDON	43		37	7.71	0.89		
24	ABANDONDED	1		1	0.21			
25	ABANDONED	78		68	14.17	0.87		
26	ABANDONING	19		18	3.75	0.72		
27	ABANDONMENT	16		13	2.71	0.68		
28	ABANDONS	4		4	0.83	0.43		
29	ABASEMENT	1		1	0.21			
30	ABASING	1		1	0.21			

frequency alphabetical statistics filenames notes

64,042 entries Row 1 0% T S 1

The Alphabetical listing also shows us some of the words but now they're in alphabetical order. **ABANDON** comes 43 times altogether, and in 37 of the 480 texts (less than 8%). **ABANDONED**, on the other hand, not only comes more often (78 times) but also in more texts (14% of them). They are similar in dispersion 354 (their spread through the corpus of texts). As explained here 324, dispersion is not shown for words which have been lemmatised after the word list was computed.

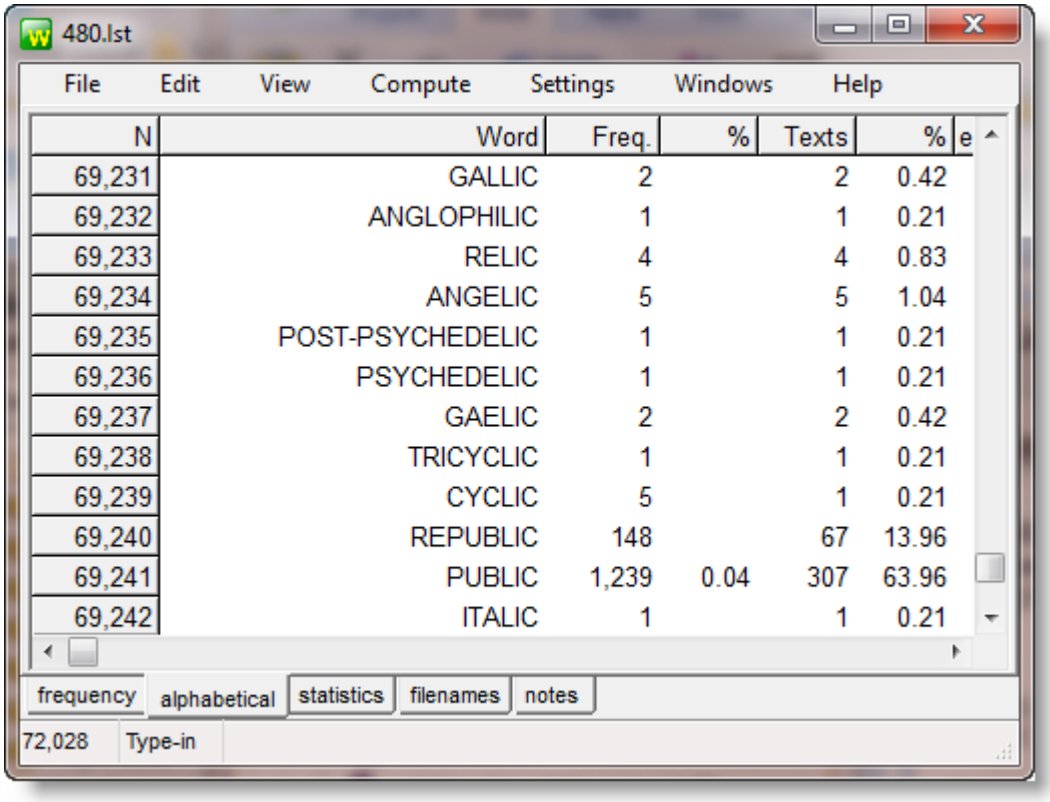
Now let's examine the statistics.



N	Overall	1	2
text file	overall	a00.txt	a01.txt
file size	16,333,680	40,079	44,749
tokens (running words) in text	2,833,815	6,820	8,041
tokens used for word list	2,833,815	6,820	8,041
sum of entries			
types (distinct words)	72,028	1,571	1,567
type/token ratio (TTR)	2.54	23.04	19.49
standardised TTR	43.65	45.03	37.70
standardised TTR std.dev.	54.68	46.98	52.28
standardised TTR basis	1,000	1,000	1,000
mean word length (in characters)	4.57	4.68	4.36
word length std.dev.	2.60	2.70	2.48
sentences	107,073	271	351
mean (in words)	26.47	25.17	22.91
std.dev.	37.59	19.56	17.24
paragraphs	34,120	110	156
mean (in words)	83.05	62.00	51.54
std dev	514.17	36.21	56.32

In all 480 texts, there are 72,028 word types (as pointed out above). The total running words is 2,833,815. Each word is about 4.57 characters in length. There are 107,073 sentences altogether, on average 26.47 words in length. In the text of **a00.txt**, there are only 1,571 different word types and that interview is under 7,000 words in length. This is explained in more detail in the Statistics 350 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**.



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

n/a and triangles

n/a and triangles in the display?

In some displays you may see **n/a** or cells with a small triangle in the corner.

BRING	3	0.28	1	100.00	n/a	bring[1] brought[2]
CUT	3	0.28	1	100.00	n/a	cut[1] cuts[2]
DO	3	0.28	1	100.00	n/a	do[0] did[2] does[1]

Here a word list was lemmatised 321 after it had been first computed. In the Dispersion column we get **n/a** because the calculation of dispersion of **bring** and **brought** takes place during the building of the word list and its value is unknown for the whole lemma.

In this key word list below, some of the statistics (e.g. BIC score 288 and Log Ratio 288) are not known because it's an old key word list computed before those statistics were available in WordSmith. And some cells have triangles to show the value isn't necessarily certain: the reference corpus total of **we've** and **we're** is uncertain after manual lemmatising 325, because it is not clear whether the reference corpus had those two words lemmatised when the key word list was computed.

YEAH	4,185	0.15	n/a	83,019	0.08	n/a	1,124.50	n/a	0.0000000000	
WE'VE	1,371	0.05	n/a	16,291	0.02	n/a	1,120.78	n/a	0.0000000000	we've[1371] we're[1439]
~~WE'RE~~	~~1,439~~	~~0.05~~	~~n/a~~	~~18,646~~	~~0.02~~	~~n/a~~	~~1,022.94~~	~~n/a~~	~~0.0000000000~~	
YESTERDAY	1,389	0.05	n/a	18,198	0.02	n/a	968.97	n/a	0.0000000000	

In the file-names below, it's because these names are from a network server. To find out whether they have been moved 122 would slow things down too much.

N	filename
2	\\DISKSTATION718\MikeStuff\text\480texts\A01.TXT
3	\\DISKSTATION718\MikeStuff\text\480texts\A0A.TXT
4	\\DISKSTATION718\MikeStuff\text\480texts\A0W.TXT
5	\\DISKSTATION718\MikeStuff\text\480texts\A13.TXT

See also : Consistency 310, Lemmatisation 322

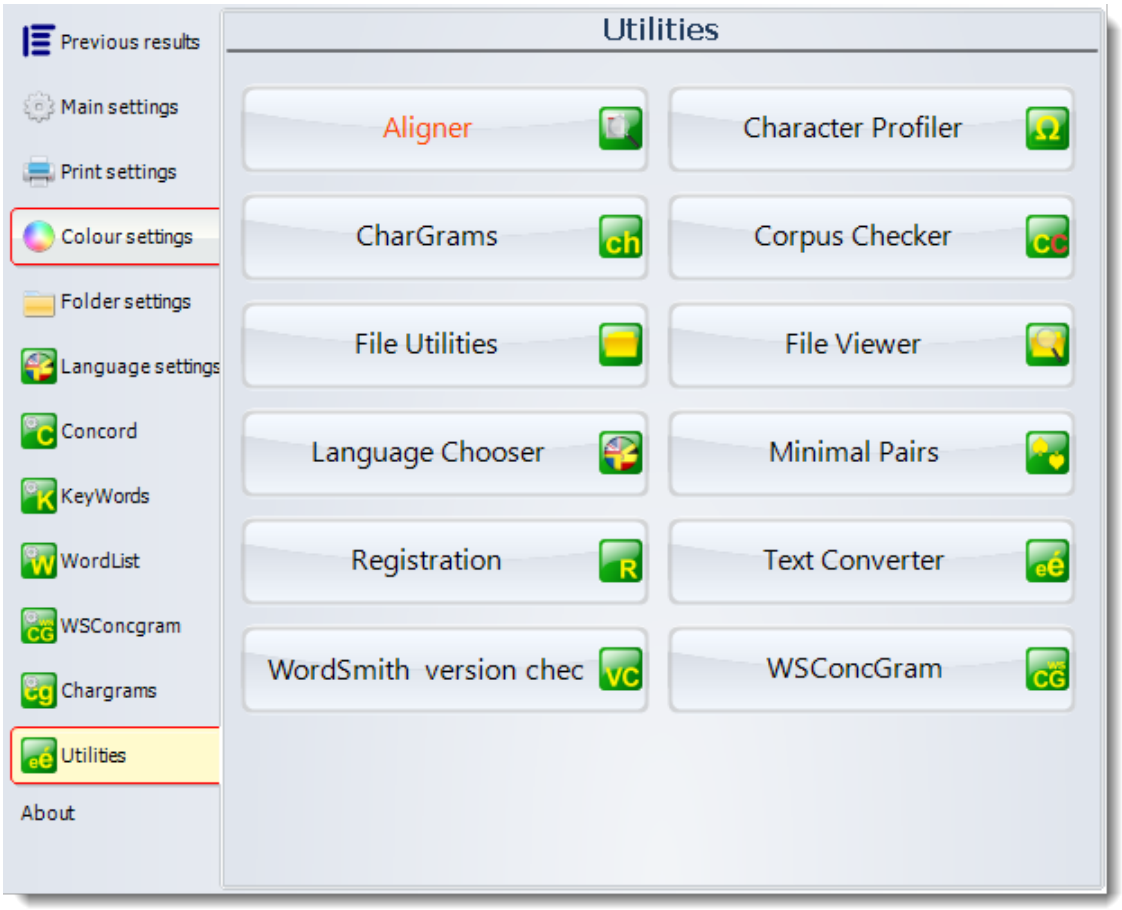
Utility Programs

10 Utility Programs

Besides the three main programs, there are more Tools that have arisen over the years; this Chapter explains them.

How to find them

You will see small icons in the main Controller under the buttons for Concord, KeyWords and WordList. Or access them via the Utilities tab:

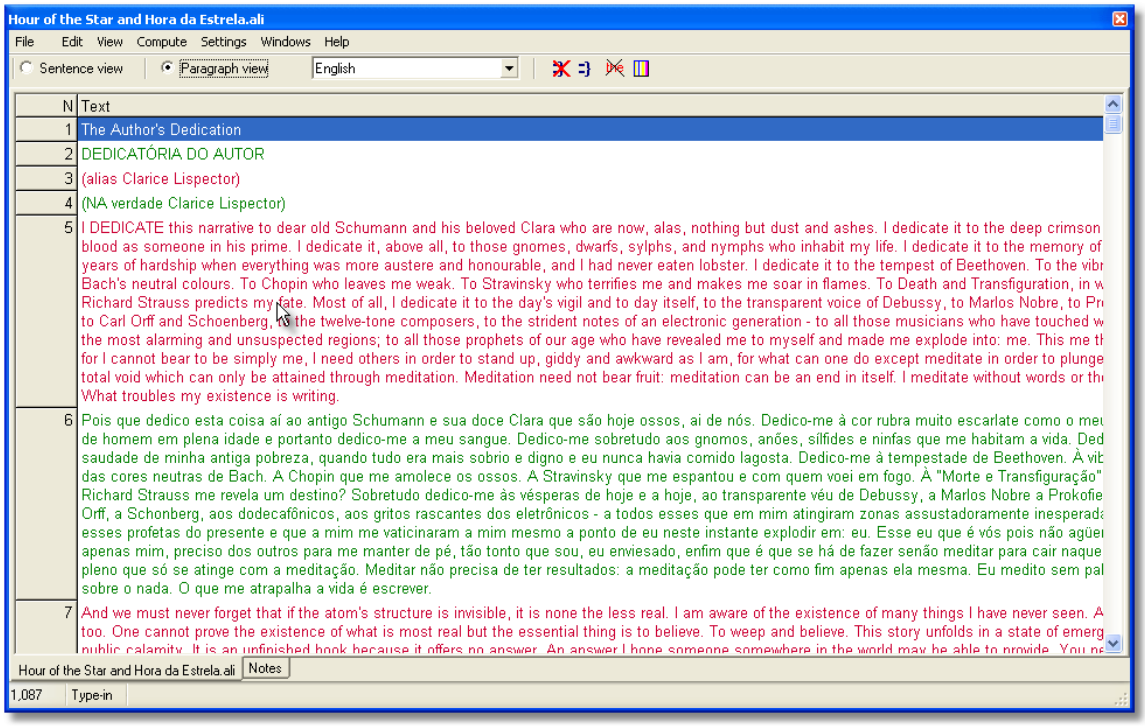


Aligner 8	helps align translated texts
Character Profiler 404	lists characters used in your texts
CharGrams 9	like WordList but for sequences of characters
Corpus Checker 10	helps to clean up your corpus, finding anomalous texts
File Utilities 10	various utilities for managing files
File Viewer 10	shows the innards of your text files
Language Chooser 9	specifies details of languages you work with
Minimal Pairs 11	identifies similar words

Registration 471	to register your licence
Text Converter 471	prepares your corpora for different formats
WordSmith version check 11	to update when there's a new build of the major version
WSConcGram 13	finds and shows concgrams

10.1 Aligner

This is a program whose main use is to produce an aligned 388 version of 2 or more texts, with alternate sentences or paragraphs from each of them.



See also: Viewer & Aligner settings 401, Viewer & Aligner options 395, an example of aligning 389

10.1.1 index

Explanations

What is the Aligner and what's it for? 387
an example of aligning 389
Settings 401
Viewing Options 395
What to do if it doesn't do what I want... 563
Searching for Short Sentences 403
Joining/Splitting 400
Aligning a Dual Text 392
Finding translation mis-matches 402
The technical side... 401

see also : WordSmith Main Index 2

10.1.2 aligning

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 77 can use it as they might use dual language readings, to study closely the differences e.g. in word order.
It helps you produce a new text which consists of the two files, with sentences interspersed. That way you can compare the translation with the original.

Example

Original : *Der Knabe sagte diesen Gedanken dem Schwesterchen, und diese folgte. Allein auch der Weg auf den Hals hinab war nicht zu finden. So klar die Sonne schien, …(from Stifter's Bergkristall, translated by Harry Steinhauer, in German Stories, Bantam Books 1961)*
Translation: *The boy communicated this thought to his sister and she followed him. But the road down the neck could not be found either. Though the sun shone clearly, ...*

Aligned text:
<G1> Der Knabe sagte diesen Gedanken dem Schwesterchen, und diese folgte.
<E1> The boy communicated this thought to his sister and she followed him.
<G2> Allein auch der Weg auf den Hals hinab war nicht zu finden.
<E2> But the road down the neck could not be found either.
<G3> So klar die Sonne schien, ...
<E3> Though the sun shone clearly, ...

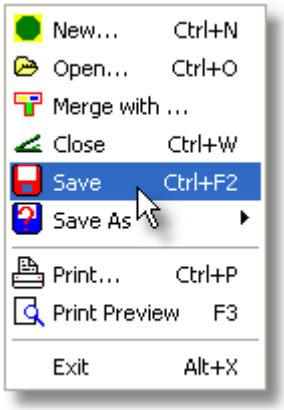
An aligned text like this helps you identify additions and omissions, normalisations, style changes, word order preferences. In this case the translator has chosen to avoid very close equivalence.

See also: an example of aligning 389, Aligning and moving 392

10.1.3 example of aligning

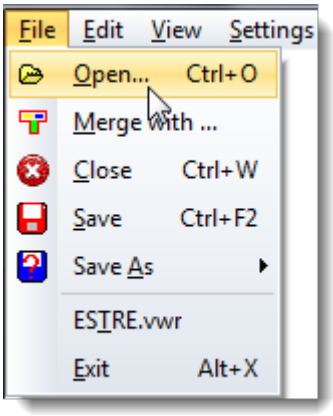
How to do it -- a Portuguese and English example

1. Read in 398 your Portuguese text (eg. **Hora da Estrela.TXT**), and checking its sentences and paragraphs break 400 the way you like. Try "Unusual Lines 403" to help identify oddities.
2. Save it

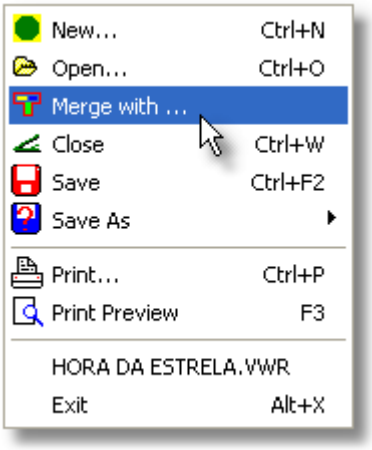


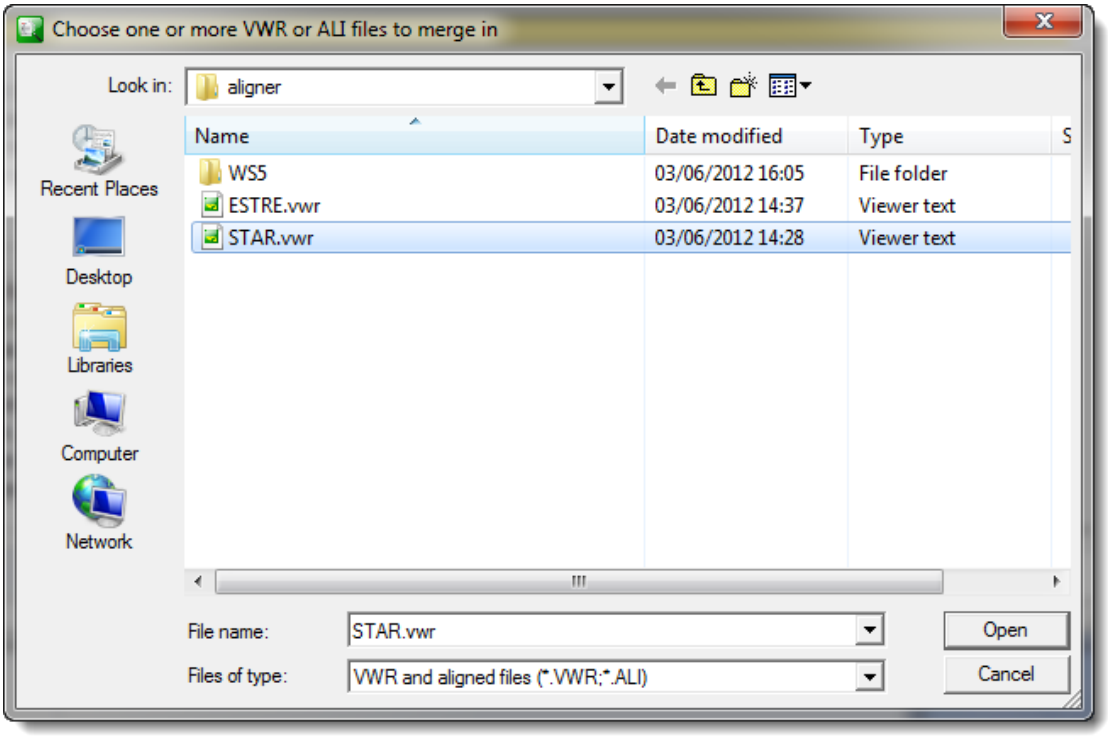
and it will (by default) get your filename **.VWR,** eg. **Hora da Estrela.VWR.**
(It is important to do that, as a .VWR file knows the language, colour settings etc. and the cleaning up work you've done, whereas the .TXT file is just the original text file you read in.)

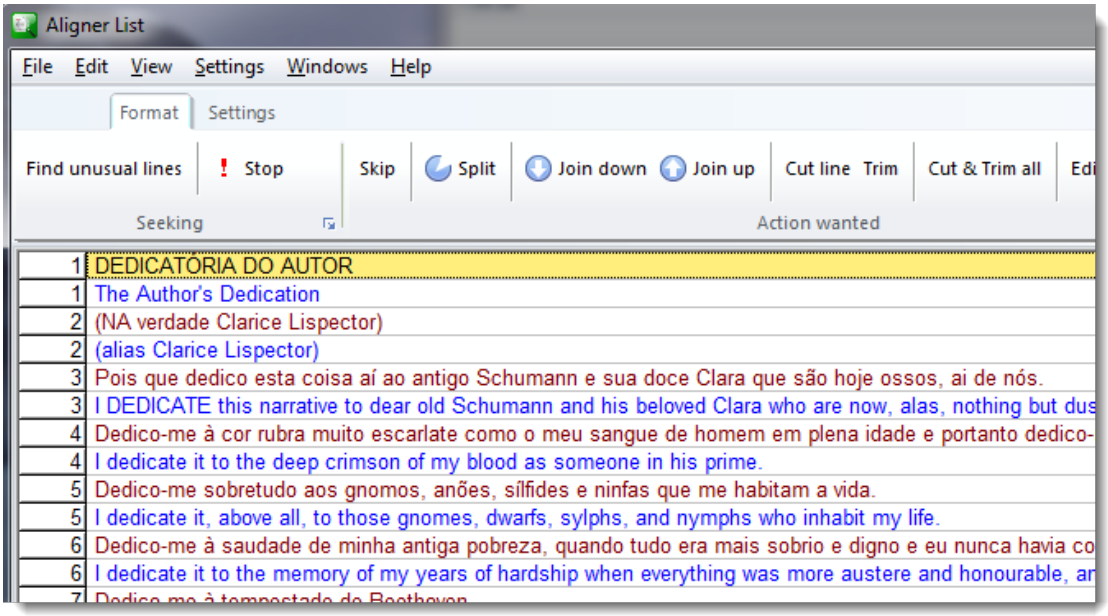
3. Do the same steps 1 and 2 for your English text -- you will now have e.g. **Hour of the Star.VWR**.
4. You could if desired repeat with the Spanish -- **Hora de la Estrella.txt** giving **Hora de la Estrella.VWR**, (or German, Russian, Arabic, etc.).
5. Now open your Portuguese **Hora da Estrela.VWR**



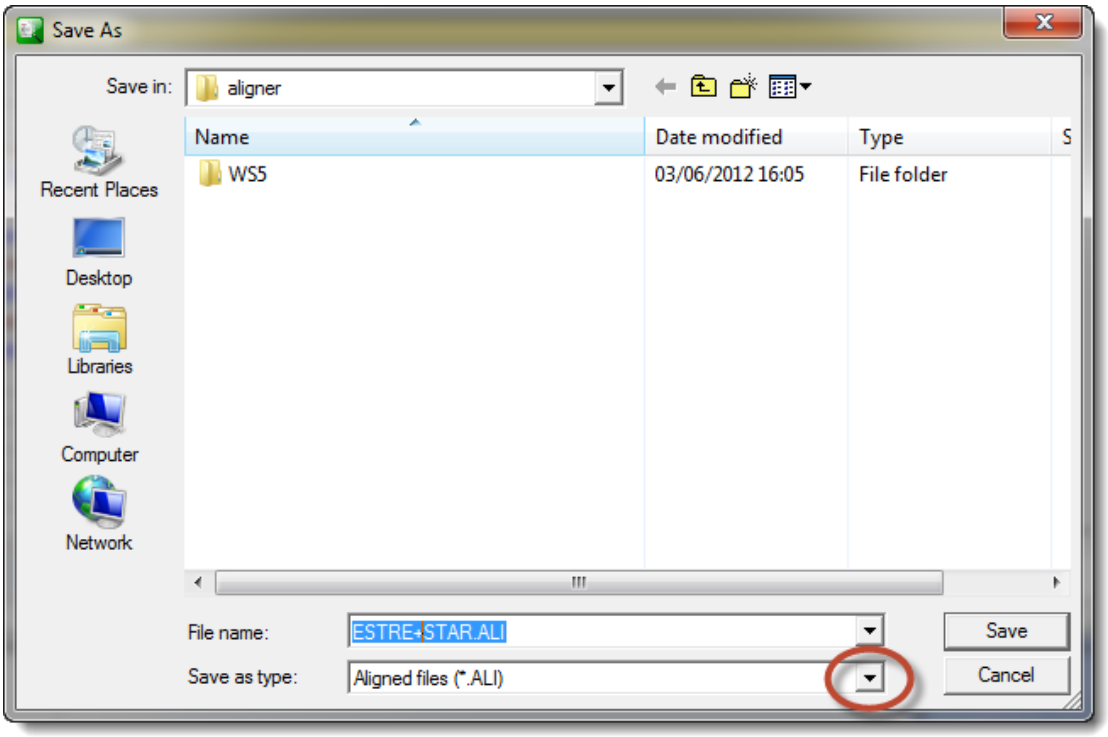
6. and then *File* | *Merge*


New... Ctrl+N
Open... Ctrl+O
Merge with ...
Close Ctrl+W
Save Ctrl+F2
Save As
Print... Ctrl+P
Print Preview F3
HORA DA ESTRELA.VWR
Exit Alt+X


Choose one or more VWR or ALI files to merge in
Look in: aligner
Recent Places
Desktop
Libraries
Computer
Network
Name
Date modified
Type
WS5
03/06/2012 16:05
File folder
ESTRE.vwr
03/06/2012 14:37
Viewer text
STAR.vwr
03/06/2012 14:28
Viewer text
File name: STAR.vwr
Files of type: VWR and aligned files (*.VWR;*.ALI)
Open
Cancel



7. Finally *File* | *Save choosing Aligned files* (**.ALI**) as the format.

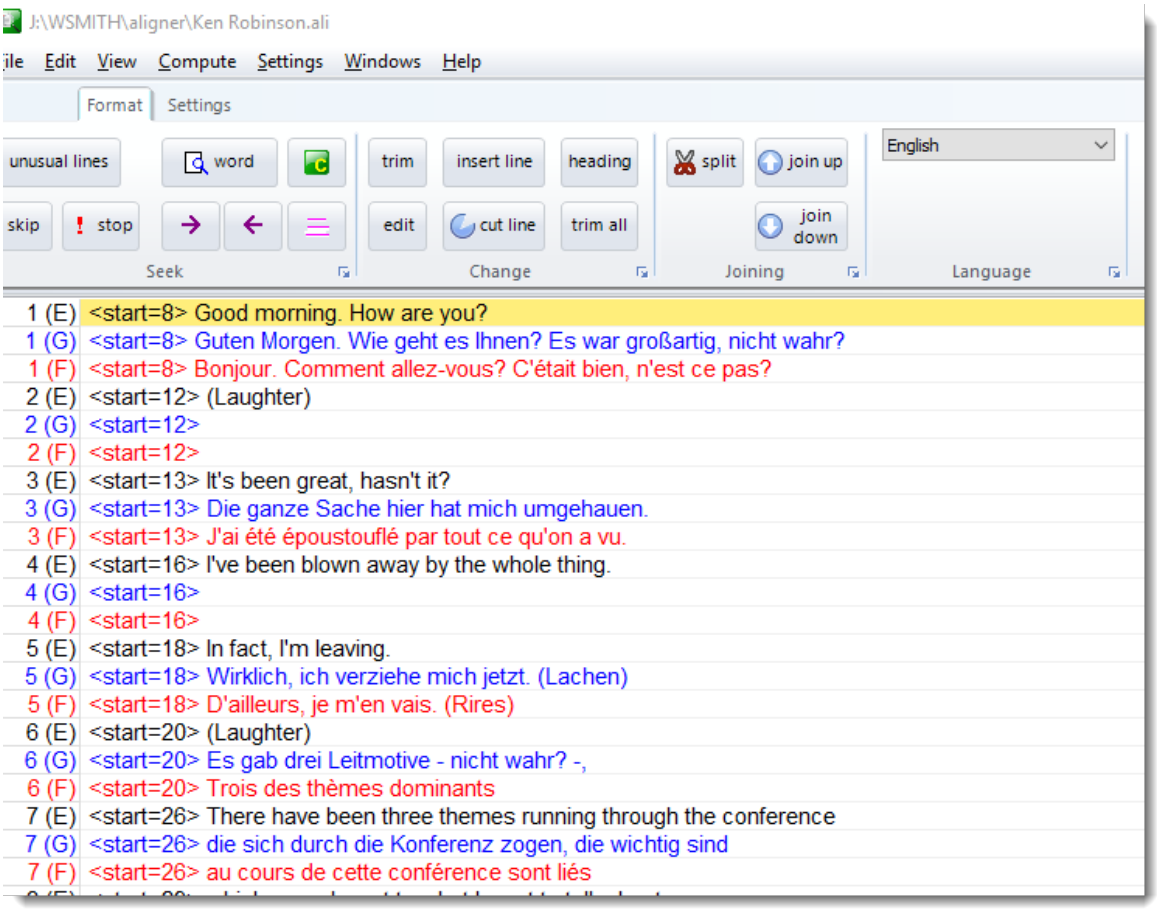


10.1.4 TED talks

You can load up transcripts of TED talks and even hear them in Aligner, just as you can in Concord 258. See obtaining sound and video files 263.

How to do it

Use File | Open to load up each transcript.



10.1.5 aligning and moving

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

adjusting by dragging with the mouse

To merge sentences or paragraphs, simply grab and drag it up to the next one above in the same language. Or use the Join button. Or press F4.
To split a sentence or paragraph, choose the Split button or press Ctrl+F4.

Finally you will want to save (Ctrl+F2) the results 91.

See also: Aligner contents 387

10.1.6 editing

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

- Split: allows you to choose where a line should be divided in two.
- Join down , Join up: these buttons merge a line with another one. You can achieve this also by simply dragging.
- Cut line: removes any blank lines.
- Trim: this goes through each sentence of the text, removing any redundant spaces -- where there are two or more consecutive spaces they will be reduced to one.
- Cut & Trim All does these actions on the whole text.
- Edit opens up a window allowing you to edit the whole of the current sentence or paragraph.
- Heading: allows you to treat a line as a heading, and if so makes it look bold.
- Find unusual lines 403: this identifies cases where a sentence or paragraph does not start with a capital letter or number -- you will probably want to join 400 it to the one above, or where a line is unusually short, etc.
- Find short lines 403

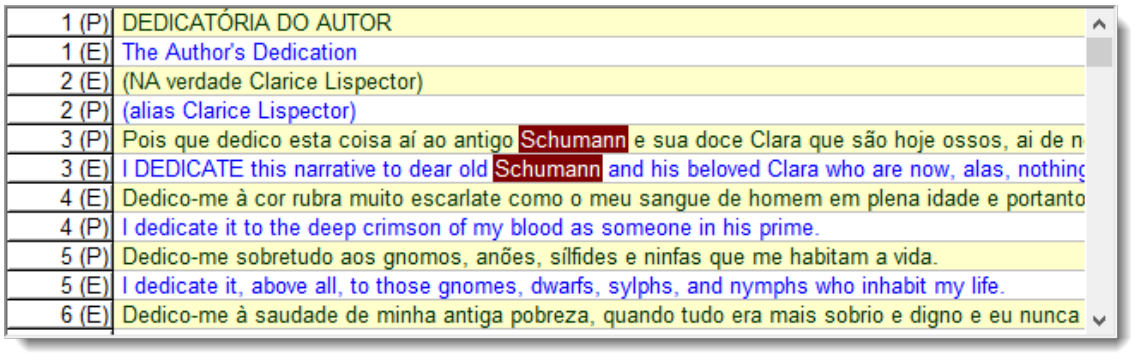
You will then want to save (Ctrl+F2) your text.

You can also:

- open a new file for viewing (you can open any number of text files within Viewer & Aligner)
- copy a text file to the clipboard 529 (select, then press Control+Ins)
- print the whole or part of the currently active text file
- search for words or phrases (press F12)

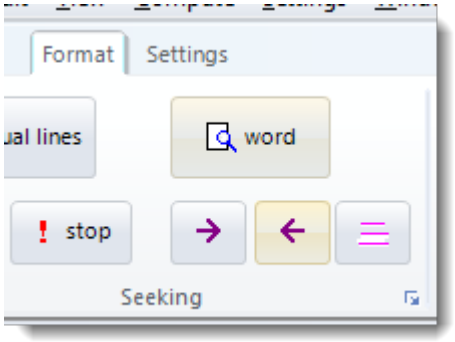
10.1.7 finding specific words

In Aligner you can search for words or phrases and get them highlighted e.g. in order to compare translations.



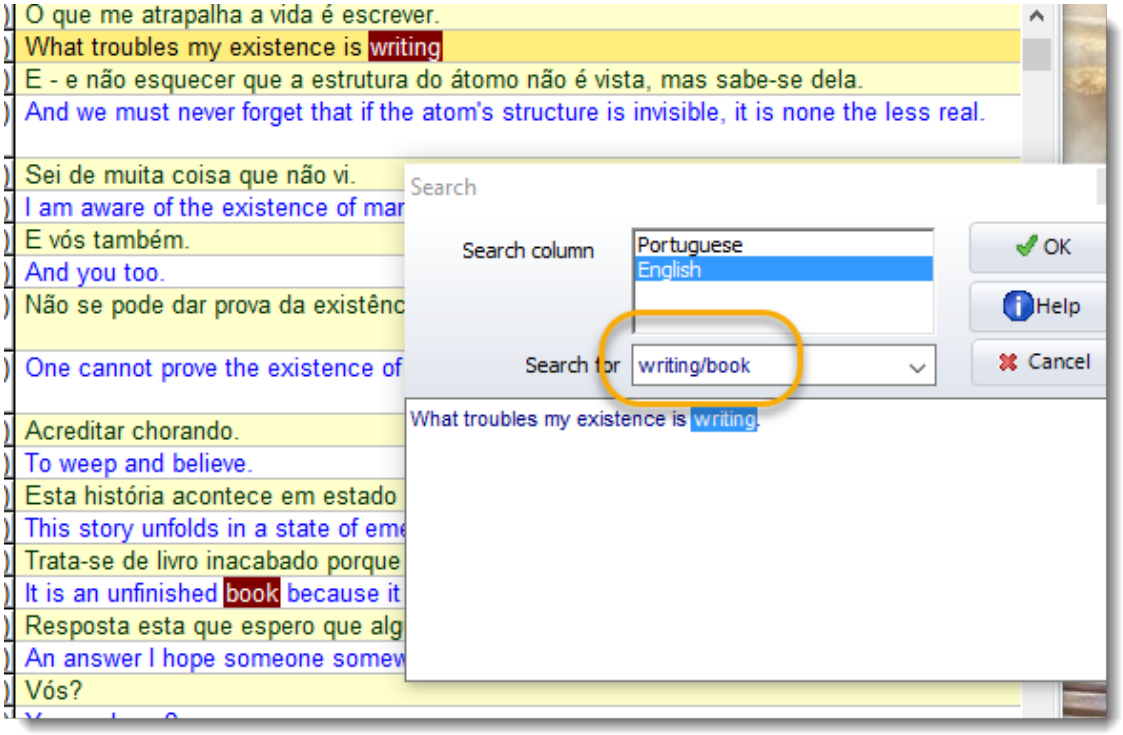
How to do it

Click the button.



Type the word or phrase you want choosing the language you want to search. If you get any hits they will be coloured, and you can press the forward ( ) and back ( ) buttons to move through the text. (Hold down Ctrl to jump 10 at a time.) You can clear the searches using the clear button ( ).

You can search on more than one word:

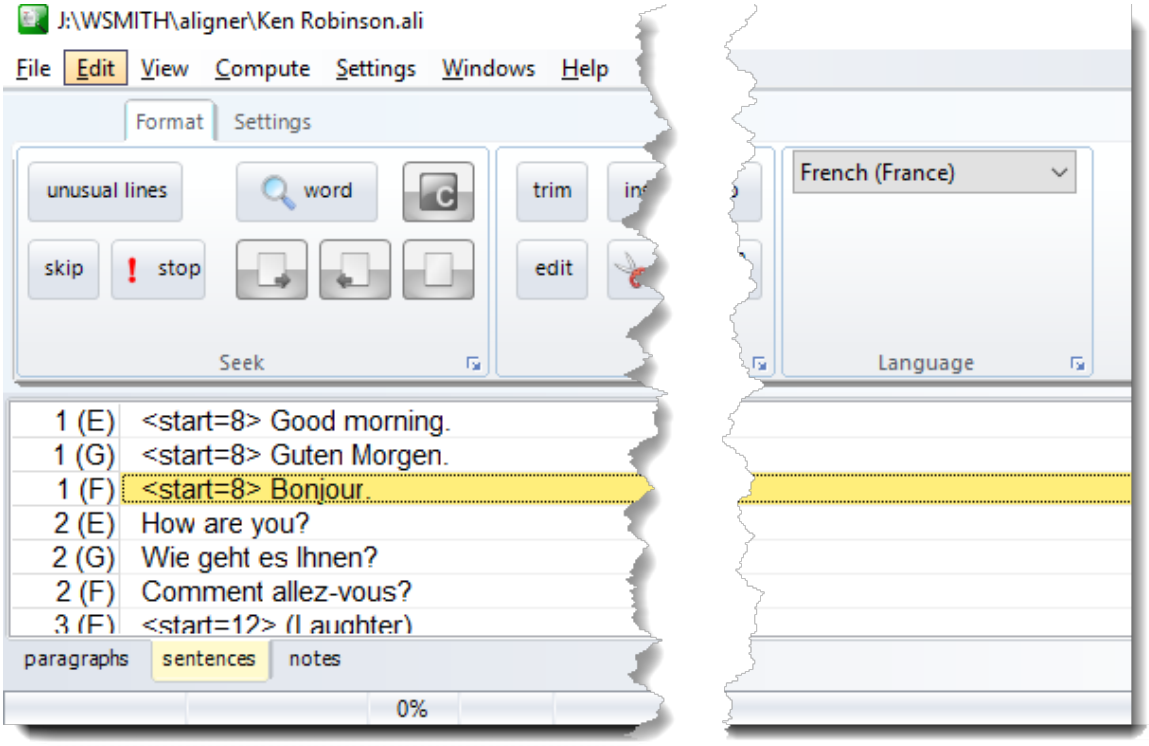


See also: search colours 401, searching 95, parallel concordancing 396

10.1.8 languages

Each **.VWR** file has its own language. Each Aligner file (**.ALI**) has one language for each of the component sections. (They could all be the same, if for example you were analysing various different editions of a Shakespeare play they'd all be English.) The set of languages available is that defined using the Languages Chooser 80.

Each time you select a line the language is shown in the language box to the right.



If you want to change the language 79, choose *Edit* | *Change language* in the menu.

10.1.9 numbering sentences & paragraphs

You can use the **Aligner** to make a copy of your text with all the sentences and/or paragraphs tagged with **<S> and <P>**.
To do this, simply read in the text file in, choose *Edit* | *Insert Tags*, then save it as a text file 91.

See also: Aligner contents 387

10.1.10 options

Mode: Sentence/Paragraph

This switches between Sentence mode and Paragraph mode. In other words you can choose to view your text files with each row of the display taking up a sentence or a paragraph.

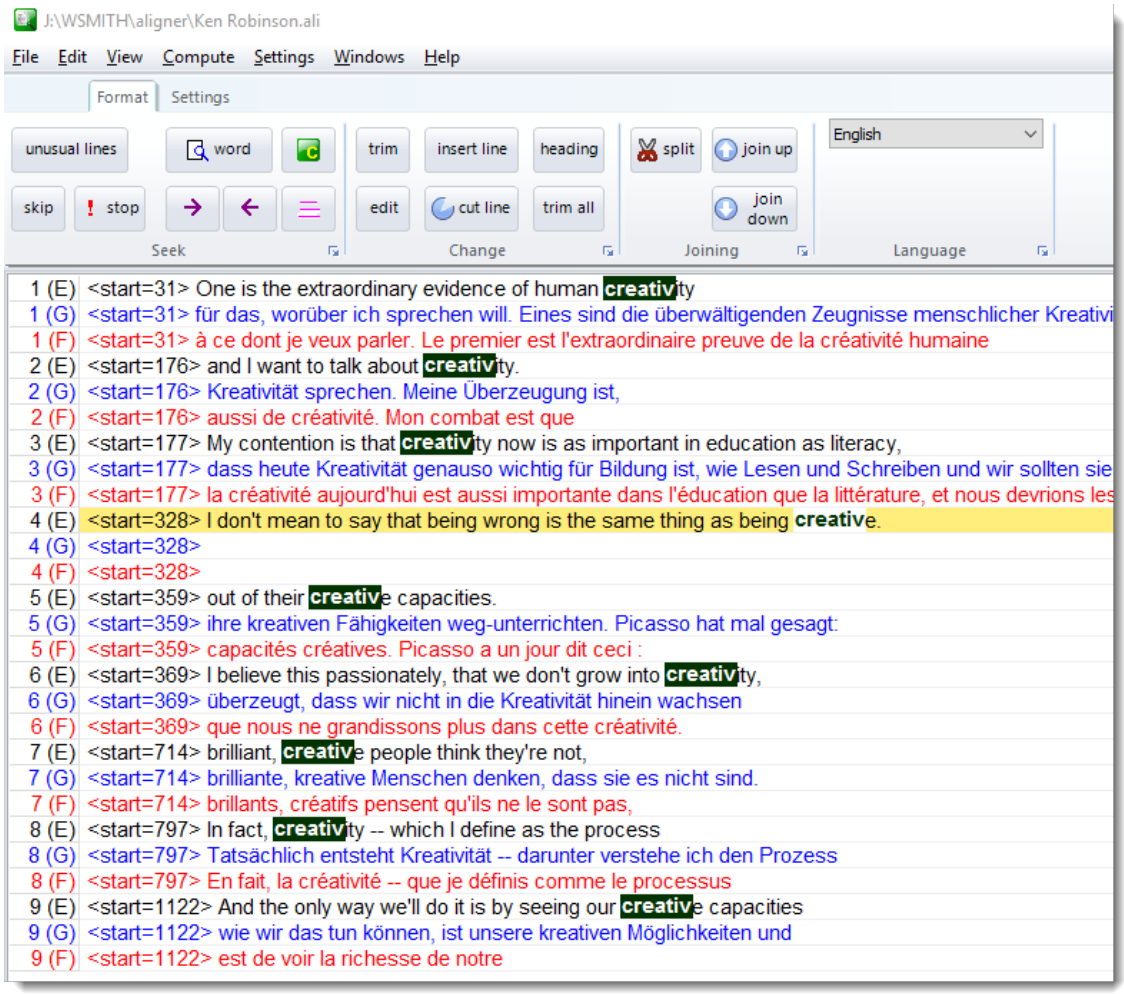
Likewise, you can make an dual aligned text by interspersing either paragraphs or sentences. The other functions (e.g. joining, splitting 400) work in the same way in either mode.

10.1.11 parallel concordancing

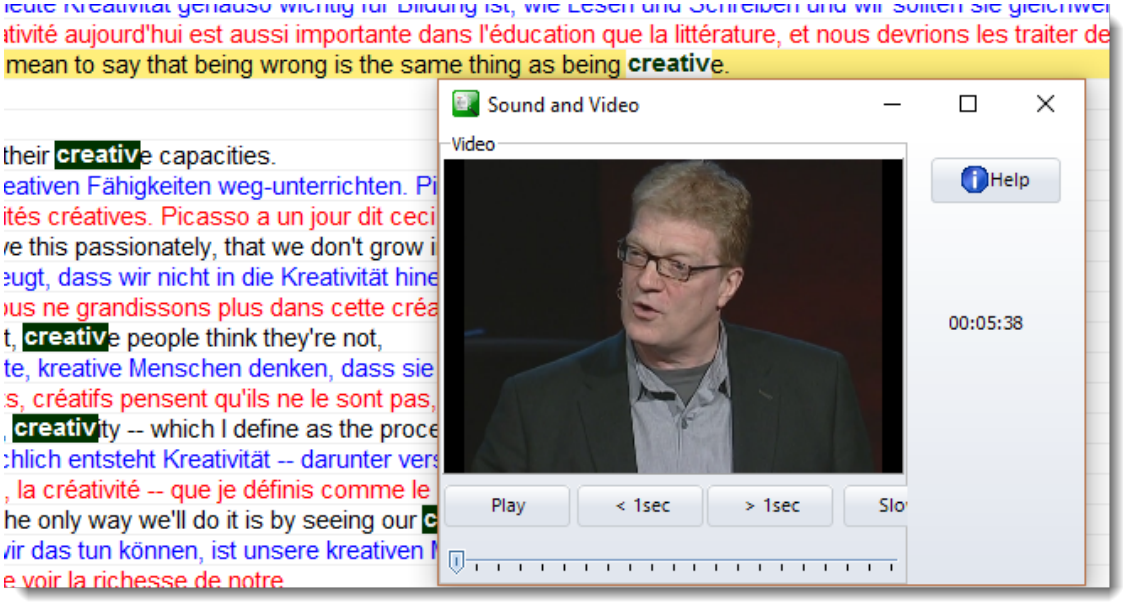
In this shot we have three versions of a TED talk 391, English, German and French.

How to do it

The word **creativ*** was searched for 393. Then all that was needed was to press the Concord ( ) button.



Only paragraphs which contain the word **creativ*** appear, along with the corresponding paragraphs in the other two languages.To hear the highlighted line, I double-clicked that line.

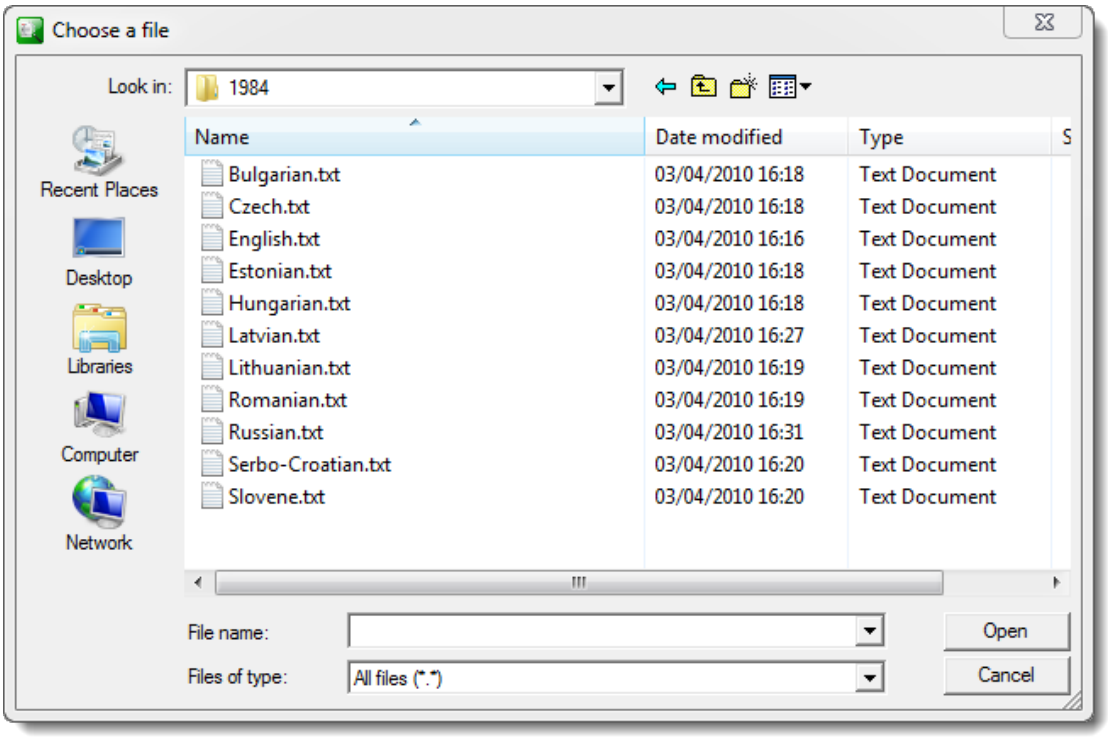


Hide the Concord tab?

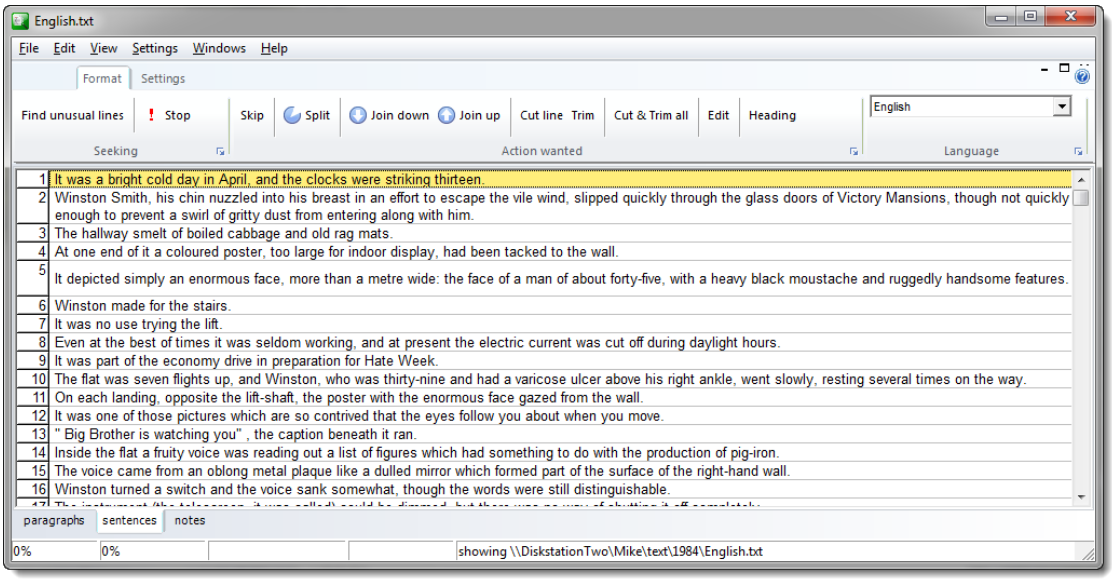
Hold down Ctrl and press the button again.

10.1.12 reading in a plain text

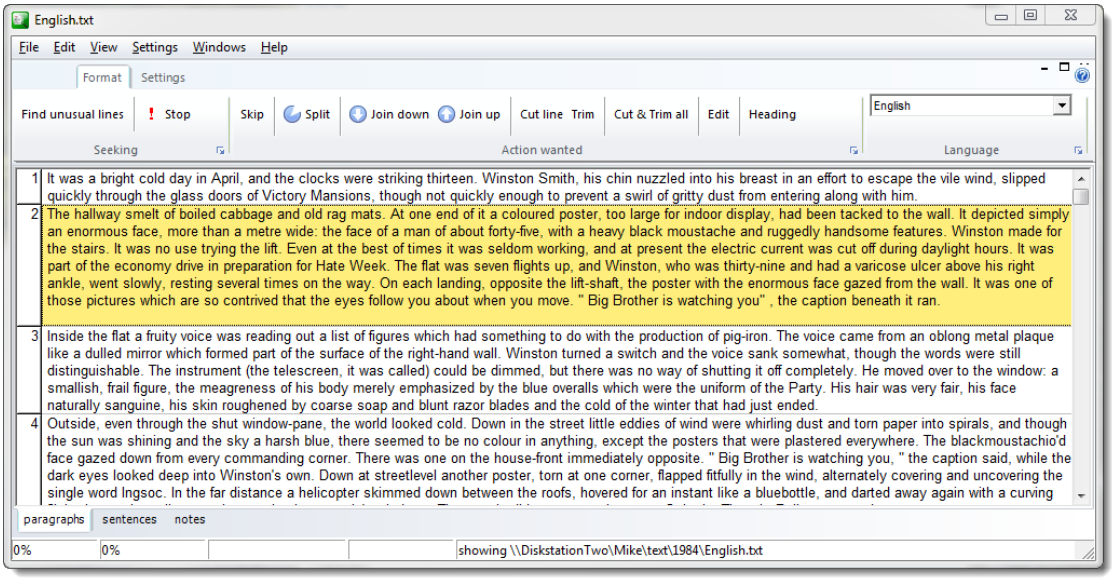
In Aligner, choose *File* | *Open,* and select your plain text file.



and you may see this sort of thing in *Sentence view*,

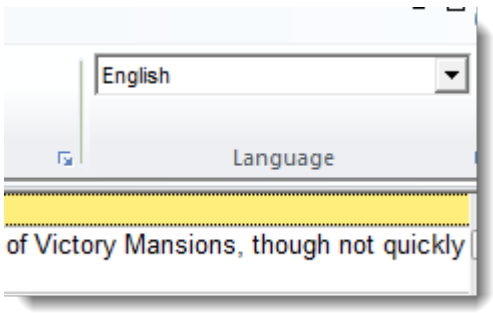


or in *Paragraph view*,

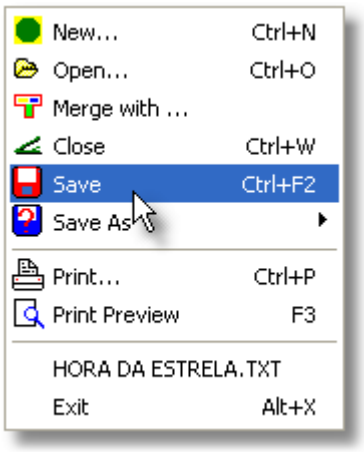


Edit it, as necessary, e.g. splitting or merging 400 paragraphs or sentences. There's a taskbar with buttons to help above the text.

Ensure the language is right:



And save it as a **.VWR** file:



See also: example of aligning 389

10.1.13 joining and splitting

Joining

The easiest way to join two sentences is simply to drag the one you want to move onto its neighbour above. Or select the lower of the two and press F4 or use the button ( )

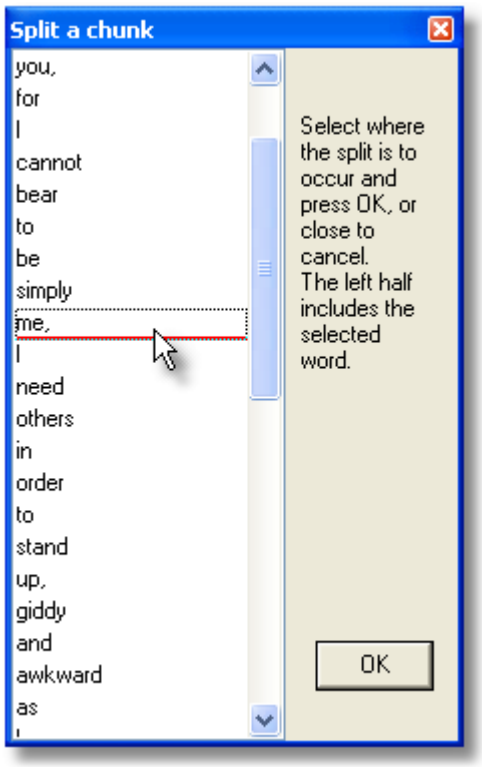
In this example, sentence 60 in Portuguese got represented as two sentences, 60 and 61, in English.

60	Que ninguém se engane, só consigo a simplicidade através de muito trabalho.
60	Let no one be mistaken.
61	Enquanto eu tiver perguntas e não houver resposta continuarei a escrever.
61	I only achieve simplicity with enormous effort.

Splitting in two

To split a sentence, press . You will get a list of the words. Click on the word which should end the sentence, then press OK.

example



This will insert the words which follow (**`I need others`** etc.) into a new line below.

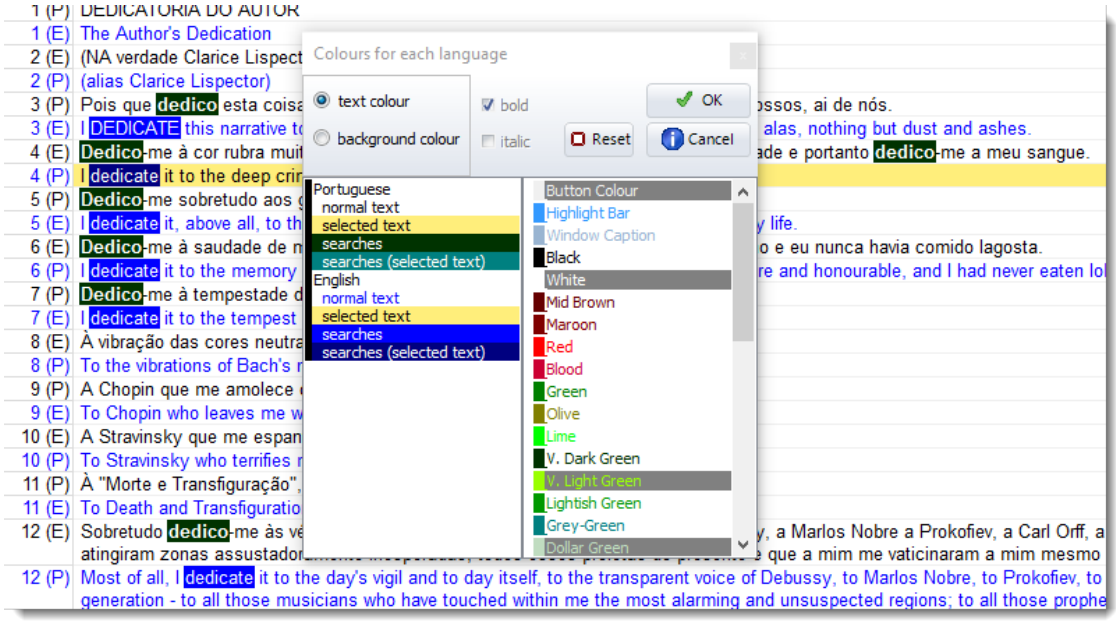
See also: Viewer & Aligner contents 387

10.1.14 settings

1. What constitutes a "short" sentence or paragraph (default: less than 25 characters)
2. Whether you want to do a lower-case check when Finding Unusual Lines

Colours

The various texts in your aligned text will have different colours associated with them. Search and ordinary display colours for each language can be changed in the Settings tab of the toolbar.



Other settings are standard ones found in most of the Tools.
Font 118
Printing 113
Text Characteristics 145
Review all Settings 110

10.1.15 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 Aligner is: a sentence ends when it meets the requirements explained in the definition of a sentence 533. The same routine is used as in WordList.

Consider this chunk from *A Tale of Two Cities*:

"Wo-ho!" said the coachman. "So, then! One more pull and you're at the top and be damned to you, for I have had trouble enough to get you to it! - Joe!"

Aligner will mistakenly consider - **`Joe!`** as a separate sentence, but handles **`"Wo-ho!" said the coachman.`** as one: though the program would split it in two if the word after **`ho!`** had a capital lettter (e.g. in **`Wild Bill, the coachman, said.`**)

Aligner cannot therefore be expected to handle all sentence boundaries exactly as *you* would. **`(I saw Mr. Smith.`** would be considered two sentences; several headings may be bundled together as one sentence.) For this reason you can choose *Find Short Sentences* to seek out 403 any odd one-word sentences.

See also: Aligner contents 387

10.1.16 translation mis-matches

Aligner can help find cases where alignment has slipped (one sentence having been translated as two or three). One method is to use the menu item *Match by Capitals*. This searches for matching proper nouns in the two versions: if say **`Paris`** is mentioned in sentences 25 of the source text and not in sentence 25 of the translation but in sentence 27, it is very likely that some slippage has occurred.

Viewer & Aligner will search forwards from the current text sentence on, and will tell you where there's a mis-match. You should then search back from that point to find where the sentences start to diverge. It may be useful to sample every 10 or every 20 to speed up the search for slippage. When you find the problem, un-join 400 or join 400 and/or edit the text as appropriate, then save it.

See also: The technical side... 401, Finding unusual sentences 403, Aligner contents 387

10.1.17 troubleshooting

Can't see the whole sentence or paragraph

Press **(autosize)** to "auto-size" the lines in your display. This adjusts line heights according to the current highlighted column of data.

Can't see the whole text file

Press to "refresh" the display.

Don't like the colours

Change colours using . The colours initially used for each language version in the dual-language window are the same colours as used for primary sorting and secondary sorting in **Concord**.

See also: Aligner contents 387

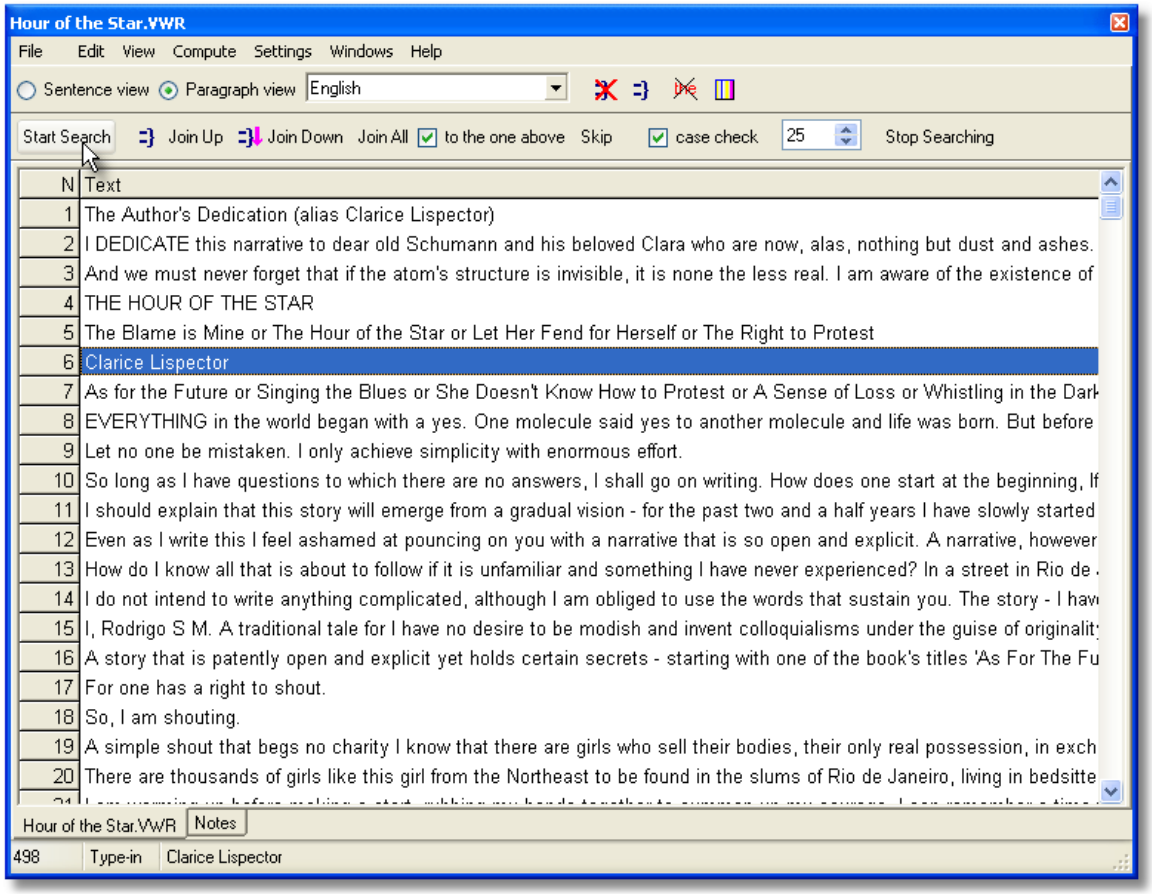
10.1.18 unusual lines

It can be useful to seek unusually short sentences to see whether your originals have been handled as you want. Because Viewer & Aligner uses full stops, question marks and exclamation marks as sentence-boundary indicators, you will find a string like "Hello! Paul! Come here!" is broken into 3 very short sentences. Depending on your purposes you may wish to consider these as one sentence, e.g. if a translator has translated them as one ("Oi, Paulo, venha cá!") .

This function can also find lower-case lines: where a sentence or paragraph does not start with a capital letter or number -- you will probably want to join it to the one above. This problem is common if the text has been saved as "text only with line breaks" (where an <Enter> comes at the end of each line whether or not it is the end of a paragraph.)

Seeking

Use the Find Unusual Toolbar menu item and then press *Start Search*. Viewer & Aligner will go to the next possibly problematic sentence or paragraph and you will probably want to join 400 it by pressing Join Up (to the one above), Join Down, or Skip.



"Case check" switches on or off the search for lower-case sentence starts. The number (25 in the example above) is for you to determine the number of characters counting as a short sentence or paragraph.

See also: Settings 401, The technical side... 401, Finding translation mis-matches 402, Viewer & Aligner contents 387

10.2 Character Profiler

10.2.1 purpose

The point of it...

Character Profiler , a tool to help find out which characters are most frequent in a text or a set of texts.

The purpose could be to check out which characters are most frequent (e.g. in normal English text the letter E followed by T will be most frequent as shown below), or it could be to check whether your text collection contains any oddities, such as accented characters or curly apostrophes you weren't expecting.

The first 32 codes used in computer storage of text are "control characters" such as tabs, line-feeds and carriage-returns. A plain **.txt** version of a text should only contain letters, numbers, punctuation and tabs, line-feeds and carriage-returns -- if there are other symbols you do not recognise you may have a **.txt** file which is really an old WordPerfect or Word **.doc** in disguise.

It would enable you to discover the most used characters across languages, as in this screenshot:

Top 10 characters										
Bulgaria	Czech	English	Estonian	Hungari	Latvian	Lithuani	Romania	Russian	Serbo-C	Slovene
a	o	e	a	e	a	i	e	o	a	e
e	e	t	e	t	i	a	a	e	o	a
o	a	a	i	a	s	s	i	a	e	i
и	n	o	s	l	e	e	r	н	i	o
н	l	i	t	n	t	t	n	и	n	n
т	t	n	l	s	n	u	u	т	s	l
с	s	h	u	k	u	o	t	л	r	r
р	v	s	k	o	r	n	c	с	t	s
в	i	r	n	i	k	r	s	р	j	j
к	k	d	d	r	l	k	l	в	u	t

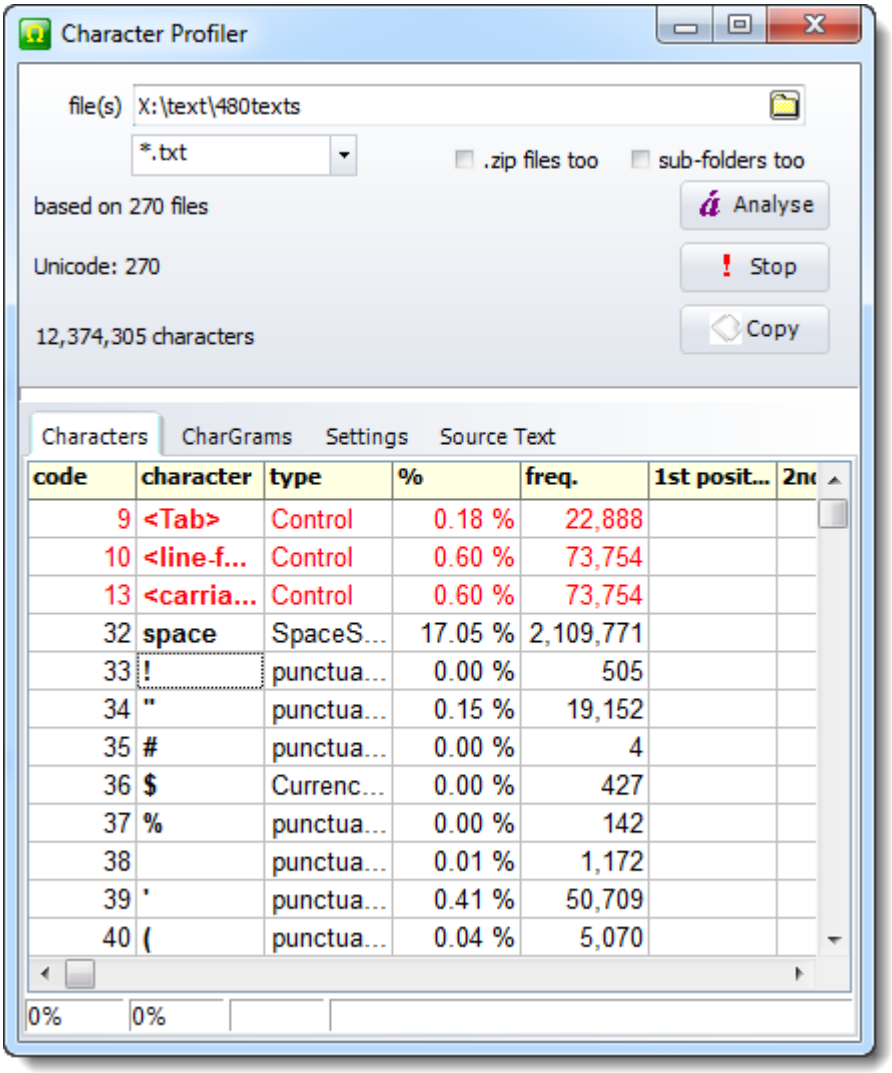
For further details see https://lexically.net/downloads/corpus_linguistics/1984_characters.xls.

See also: File Viewer 10, characters within a word 146

10.2.2 profiling characters

How to do it

1. Choose one or more texts or a folder. You can type in a complete filename (including drive and folder), and can use wildcards such as **`*.txt`**, or you can browse to find your text or folder.
2. If you want to study one text only, just choose one text, but you may choose a whole folderful or more by using the "sub-folders too" option.
3. Press *Analyse*.



The display shows details of your selected text, and if you click the *Source Text* tab you can see the original text. (If you have analysed a whole set of text files the *Source Text* tab will show only that last one.)

Legend

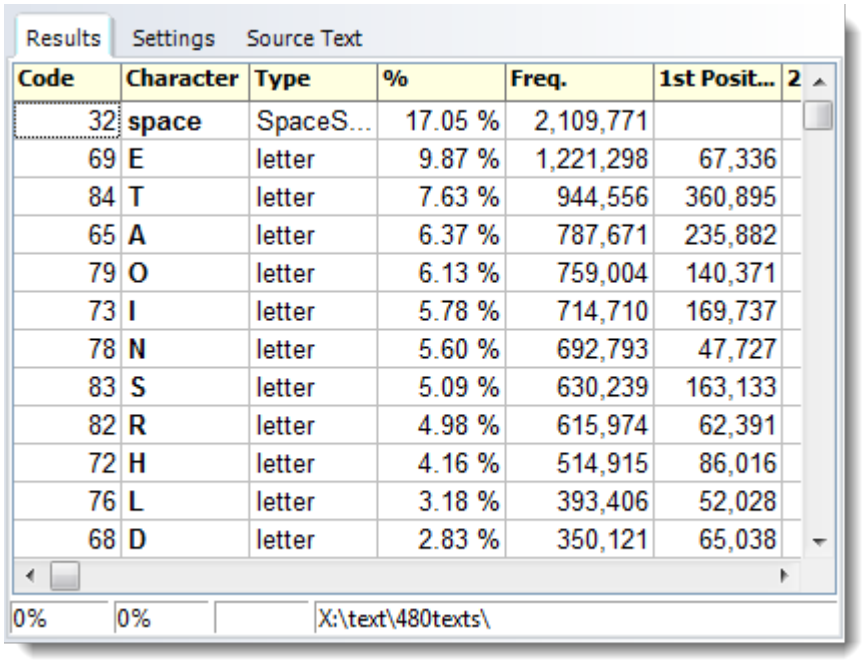
code	the Unicode code of

character	the character
type	distinguishing punctuation, digits, letters
%	percentage of the total number of characters in the text(s)
freq.	number of occurrences of that character
<Tab> etc.	control characters indicated in red
1st Position	number of each letter-character occurring in word-initial position
2nd	number found in second position in any word
etc.	

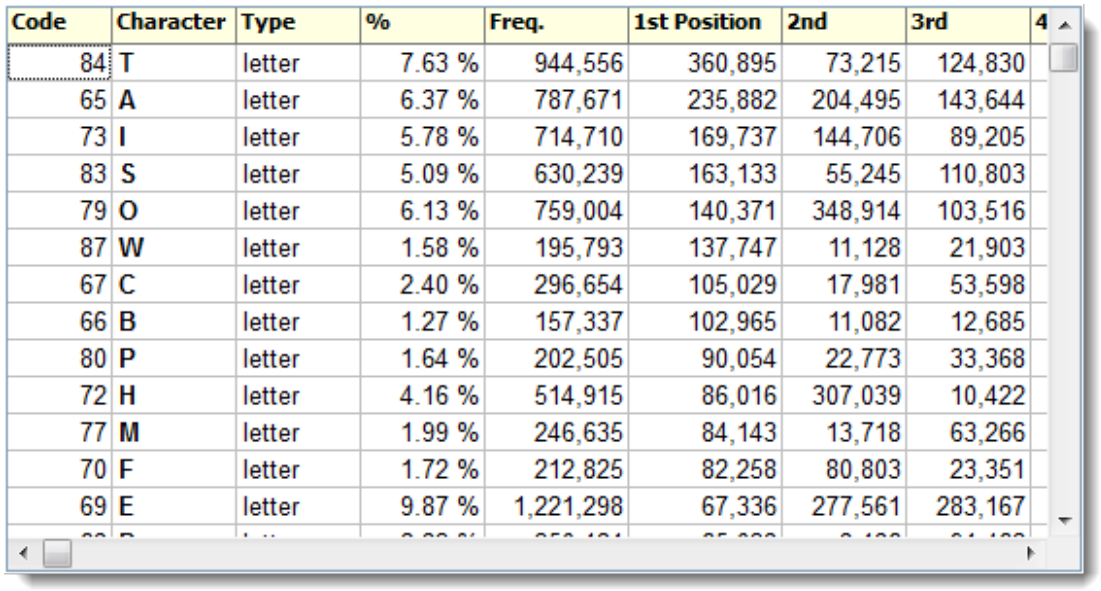
Note that 8th will only be able to count letter frequencies for words at least 8 letters long, while 1st or 2nd will handle nearly all words.

Sort

Click the header to sort the data:



The letter **E** (upper and lower case merged) here represents nearly ten percent of all letters, closely followed by **T**. If sorted by 1st position in the word, however,



Code	Character	Type	%	Freq.	1st Position	2nd	3rd	4
84	T	letter	7.63 %	944,556	360,895	73,215	124,830	
65	A	letter	6.37 %	787,671	235,882	204,495	143,644	
73	I	letter	5.78 %	714,710	169,737	144,706	89,205	
83	S	letter	5.09 %	630,239	163,133	55,245	110,803	
79	O	letter	6.13 %	759,004	140,371	348,914	103,516	
87	W	letter	1.58 %	195,793	137,747	11,128	21,903	
67	C	letter	2.40 %	296,654	105,029	17,981	53,598	
66	B	letter	1.27 %	157,337	102,965	11,082	12,685	
80	P	letter	1.64 %	202,505	90,054	22,773	33,368	
72	H	letter	4.16 %	514,915	86,016	307,039	10,422	
77	M	letter	1.99 %	246,635	84,143	13,718	63,266	
70	F	letter	1.72 %	212,825	82,258	80,803	23,351	
69	E	letter	9.87 %	1,221,298	67,336	277,561	283,167	

the letter **E** comes after **T,A,I,S,O,W,C,B,P,H,M** and **F** in frequency. Presumably the ranking of **T** reflects the frequency in English of **the** and **A** of **a**.

Copy

Copies the data to the clipboard, ready to be pasted for example into Excel.

See also: settings 410

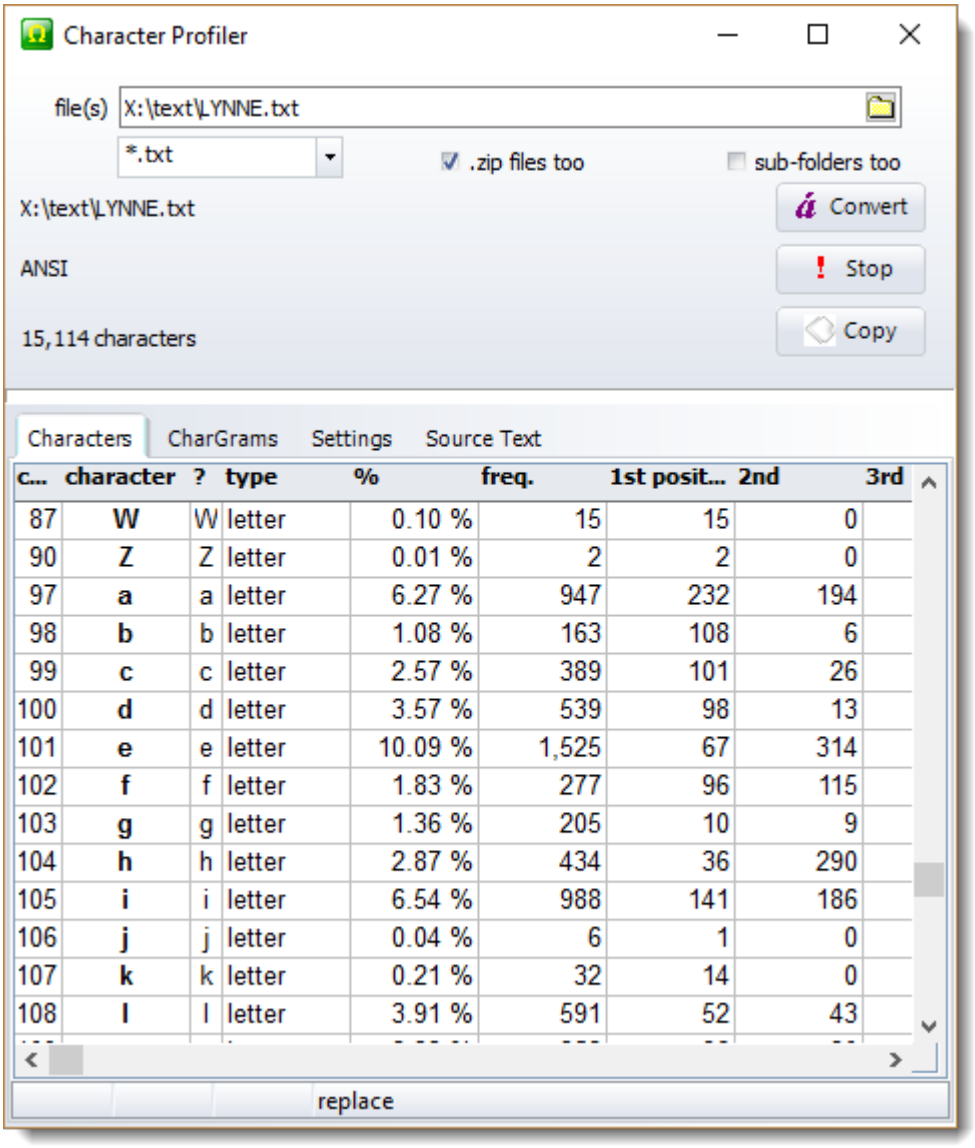
10.2.2.1 converting

The point of it

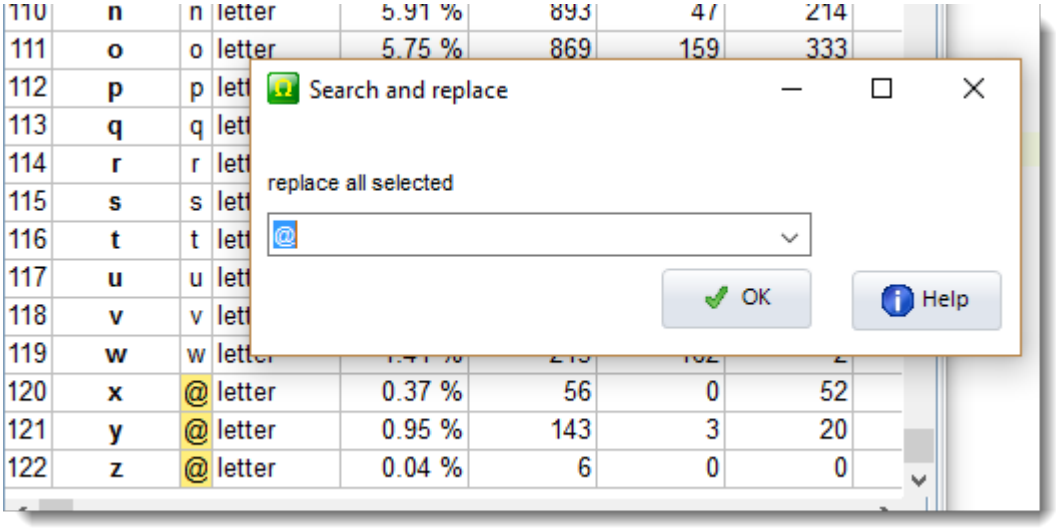
Text downloaded can often contain strange or unwanted characters and this function allows you to replace these.

How to do it

After you have analysed 405 your text, the *Analyse* button text changes to the word *Convert*.



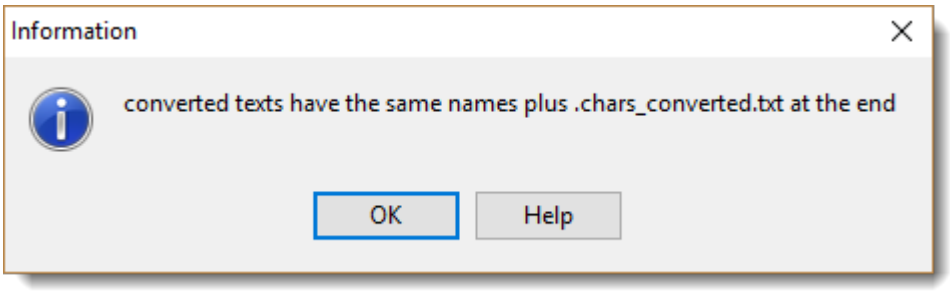
You can now select any of the characters in the 3rd column headed ?, and press *replace* in the status bar below. In the screen shot below I selected **x**, **y**,and **z** and chose to replace them with **@**.



Then by selecting **s** and pressing *replace* again, I chose to replace **s** with #.

Finally, press Convert.

If you give the go-ahead, your text files will be re-processed and every character you wish to replace will get replaced.

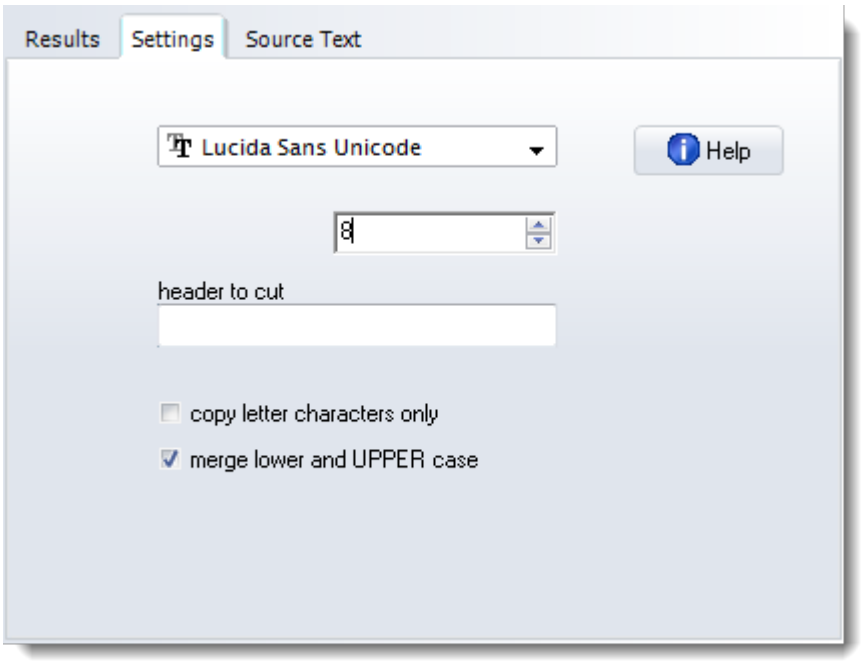


The result (in this test case) all looked like this fragment:

```
        In 1992 Hong Kong Government commi##ioned Con#ultant# to look at the
infra#tructure requirement# for an i#land e@ten#ion to the Hong Kong Convention
and E@hibition Centre (HKCEC).  In the #ame @ear, the Environmental Protection
Department carried out an apprai#al of environmental i##ue# relating to the
#cheme.  A number of concern# aro#e from thi# apprai#al and in order that the#e
be properl@ addre##ed, a multi-di#ciplinar@ focu#ed #tud@ wa# commi##ioned.(p)
```

See also: curly quote marks 491, converting using Text Converter 481

10.2.3 profiling settings



The top two boxes allow you to choose a font for your display. Most fonts can only represent some of the Unicode characters, so you may need to experiment to determine which is best for your language. (Character Profiler translates any text into Unicode whether or not it is in Unicode originally, and tells you which form it is in on the Results tab.)

Header to cut

If you've typed something in here such as **</Header>**, the program treats all the text before that as a header to be excluded from analysis..

Copy letter characters only

Check this one to force the copying to the clipboard to copy only data of letters, ignoring punctuation and digits.

Merge lower and UPPER case

Check this one to convert all text to upper case.

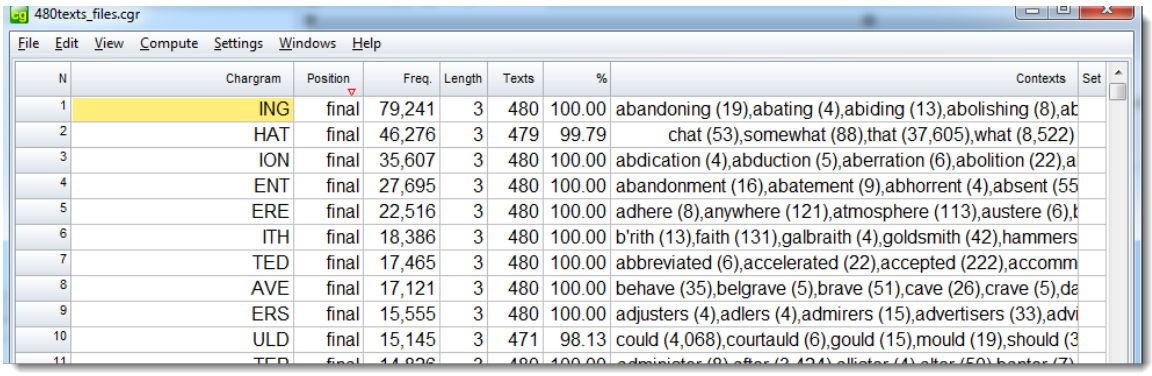
10.3 Chargrams

10.3.1 purpose

The point of it...

Chargrams , a tool to help find out which chargrams 534 (sequences of N characters) are most frequent in a text or a set of texts.

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



N	Chargram	Position	Freq.	Length	Texts	%	Contexts	Set
1	ING	final	79,241	3	480	100.00	abandoning (19),abating (4),abiding (13),abolishing (8),ab	
2	HAT	final	46,276	3	479	99.79	chat (53),somewhat (88),that (37,605),what (8,522)	
3	ION	final	35,607	3	480	100.00	abdication (4),abduction (5),aberration (6),abolition (22),a	
4	ENT	final	27,695	3	480	100.00	abandonment (16),abatement (9),abhorrent (4),absent (55	
5	ERE	final	22,516	3	480	100.00	adhere (8),anywhere (121),atmosphere (113),austere (6),b	
6	ITH	final	18,386	3	480	100.00	b'rith (13),faith (131),galbraith (4),goldsmith (42),hammers	
7	TED	final	17,465	3	480	100.00	abbreviated (6),accelerated (22),accepted (222),accomm	
8	AVE	final	17,121	3	480	100.00	behave (35),belgrave (5),brave (51),cave (26),crave (5),da	
9	ERS	final	15,555	3	480	100.00	adjusters (4),adlers (4),admirers (15),advertisers (33),advi	
10	ULD	final	15,145	3	471	98.13	could (4,068),courtauld (6),gould (15),mould (19),should (3	

These are 3-letter chargrams occurring in word-final position. **ING** is a well-known ending in English; **HAT** is is a frequent 3-letter sequence at the end of words too.

How does it work?

Chargrams are computed by taking only the valid characters of text. If a text contained "In 1845 there was a princess", the 3-character chargrams considered would be **THE, HER, ERE, WAS, PRI, RIN, INC, NCE, CES, ESS**. The positions are computed in relation to the original words, so **THE** is word-initial while **ESS** is word-final, and **RIN** is medial.

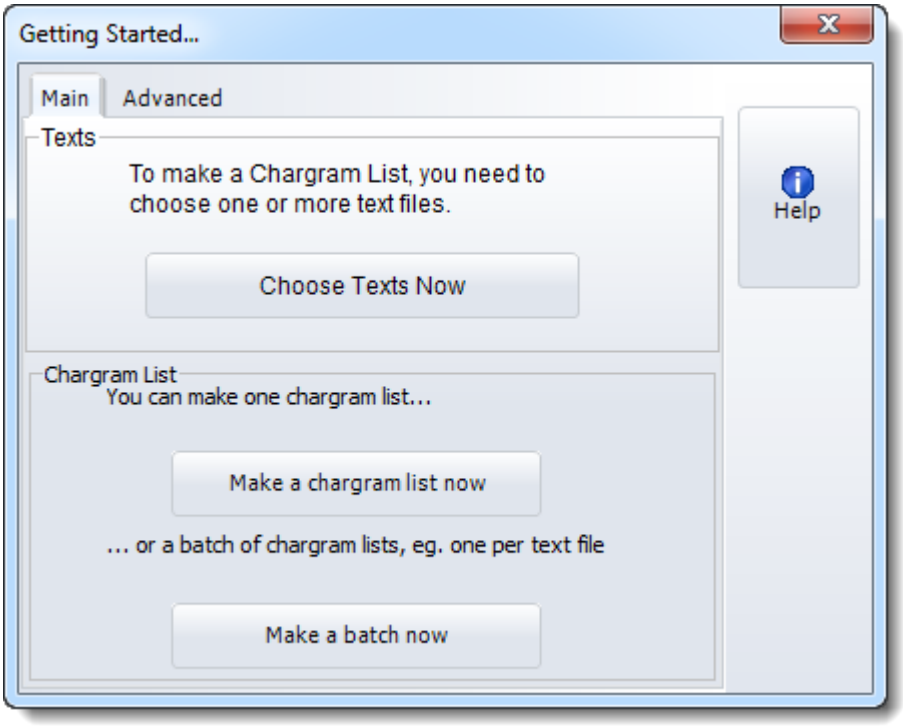
If including punctuation, the sequences would include **IN_, N_1, _18, 184** etc. too.

10.3.2 chargram procedure

How to do it

Choose *File* | *New* in the Chargrams menu.

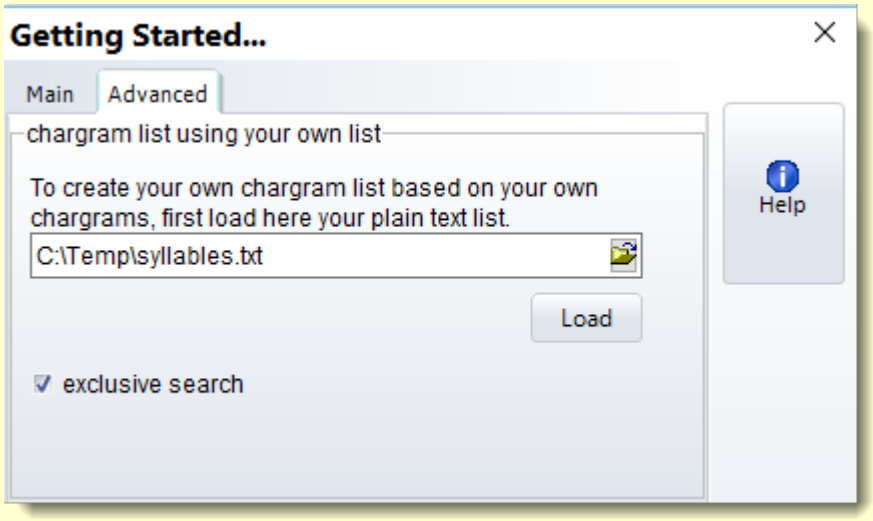
Choose your texts as with the other Tools.



Then press the button to make a chargram list.

Advanced

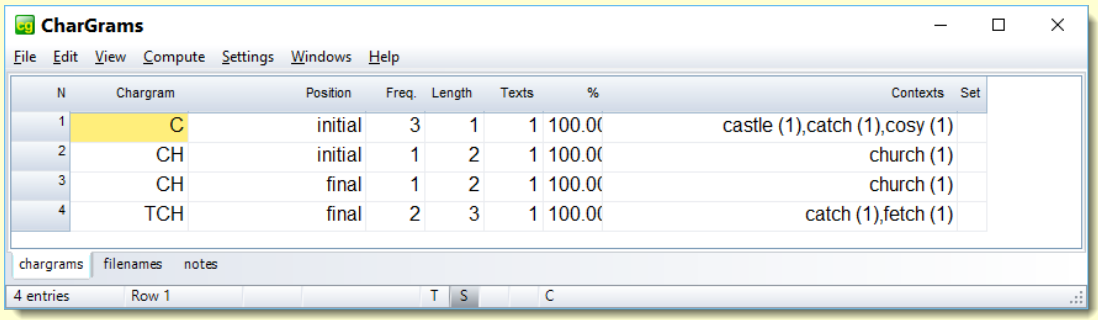
You can also use your own defined list of potential chargrams or syllables, depending on your needs.



Here we have selected a list of possible syllables in a test file (the list contained only **c, ch, tch**) and a text file consisting of a few simple words: **elephant cosy castle church catch fetch the pigs.**

Exclusive search: with this setting,for example if **tch** is found in **catch**, then no other chargrams in the list (e.g. **c, ch**) may also be counted in that context.

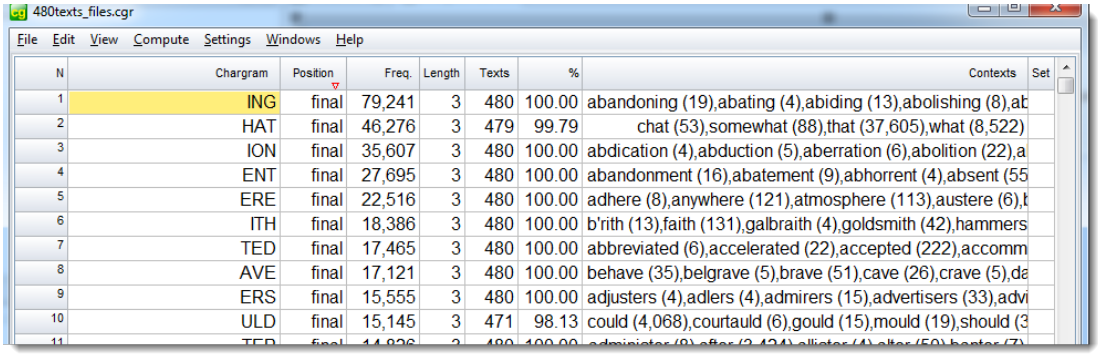
This gave the following result:



See also : settings 418.

10.3.3 display

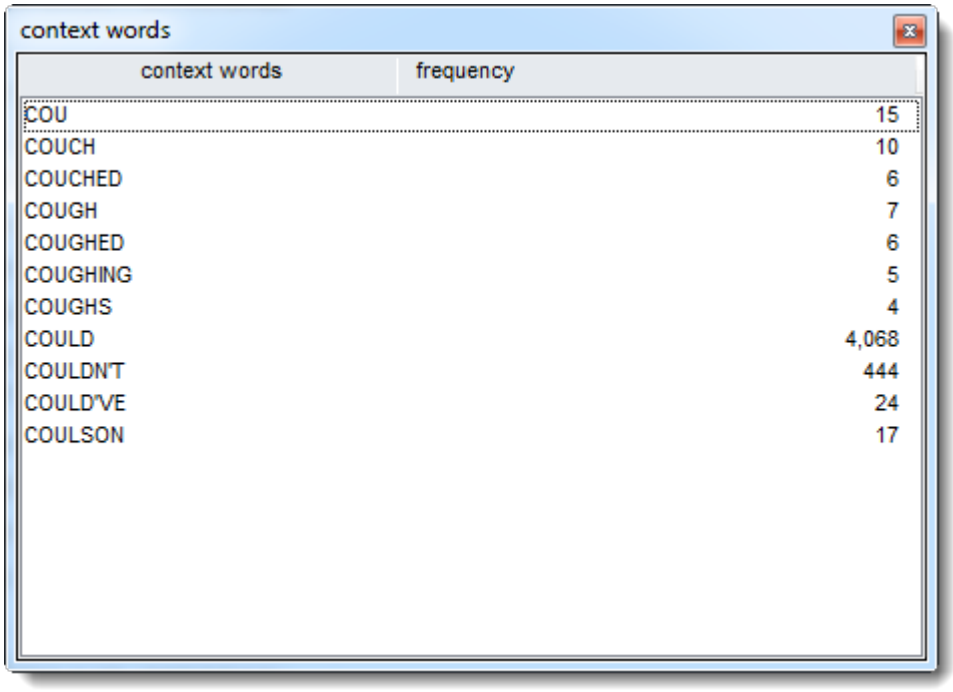
The display is similar to that in the WordList tool.



N	Chargram	Position	Freq.	Length	Texts	%	Contexts	Set
1	ING	final	79,241	3	480	100.00	abandoning (19),abating (4),abiding (13),abolishing (8),ab	
2	HAT	final	46,276	3	479	99.79	chat (53),somewhat (88),that (37,605),what (8,522)	
3	ION	final	35,607	3	480	100.00	abdication (4),abduction (5),aberration (6),abolition (22),a	
4	ENT	final	27,695	3	480	100.00	abandonment (16),abatement (9),abhorrent (4),absent (55	
5	ERE	final	22,516	3	480	100.00	adhere (8),anywhere (121),atmosphere (113),austere (6),b	
6	ITH	final	18,386	3	480	100.00	b'rith (13),faith (131),galbraith (4),goldsmith (42),hammers	
7	TED	final	17,465	3	480	100.00	abbreviated (6),accelerated (22),accepted (222),accomm	
8	AVE	final	17,121	3	480	100.00	behave (35),belgrave (5),brave (51),cave (26),crave (5),da	
9	ERS	final	15,555	3	480	100.00	adjusters (4),adlers (4),admirers (15),advertisers (33),advi	
10	ULD	final	15,145	3	471	98.13	could (4,068),courtauld (6),gould (15),mould (19),should (3	

Contexts

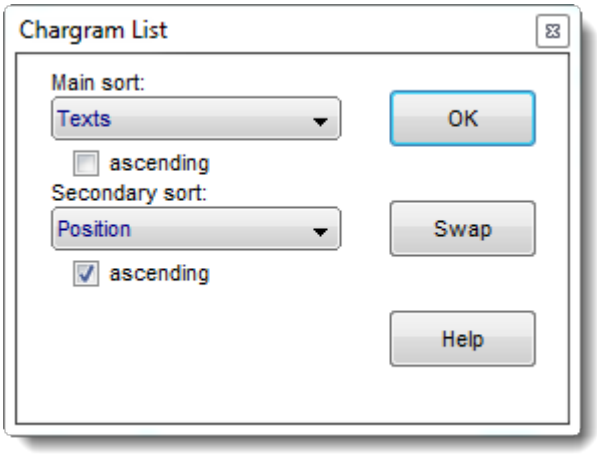
This column shows the word-contexts for each chargram. You can double-click to see the whole list.



context words	frequency
COU	15
COUCH	10
COUCHED	6
COUGH	7
COUGHED	6
COUGHING	5
COUGHS	4
COULD	4,068
COULDN'T	444
COULD'VE	24
COULSON	17

Sorting

You can sort by clicking a header. This offers you two of the columns to sort on, a primary sort and then where values on the primary sort are the same a secondary sort. In this example the user is choosing the number of texts in descending order and then word-position in ascending order.



This gave the following list (extract):

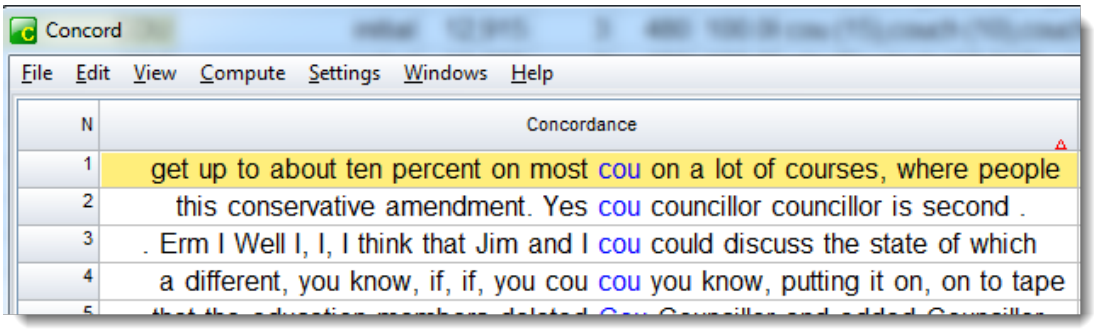
N	Chargram	Position
1	WHO	initial
2	COU	initial
3	WHI	initial
4	PRO	initial
5	THE	initial
6	BUT	initial
7	ALL	initial
8	AND	initial
9	FOR	initial
10	MOR	initial
11	THI	initial
12	CON	initial
13	WIT	initial
14	THA	initial
15	COM	initial
16	LLE	medial
17	AIN	medial
18	EAR	medial
19	ALL	medial
20	NTI	medial

Word-initial chargrams 1-15 and mid-word chargrams 16 onwards all occurred in all 100% of the texts selected.

Concordancing

As in many other Tools, you can concordance selected items by choosing *Compute* | *Concordance* in the menu.

In the case above I wondered about the context word **cou**



... which clearly shows speech reformulation.

See also : settings 418.

10.3.4 patterns

The point of it

The idea is to be able to see the neighbouring chargrams found beside a given chargram. The display is similar to a Concord patterns display 237.

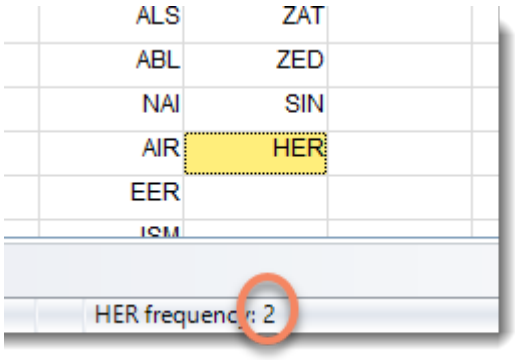
Here the chargram **ION** has been chosen. It is mostly a word final chargram in English. In L1 position we find first **NAT (nation)**, then **RAT (ration)**, **LAT** etc. To the right we get **SHI (-tionship).**

L3	L2	L1	Centre	R1	R2	R3	Set
CON	TER	NAT	ION	SHI	SAT	ION	
ORG	FOR	RAT		ALL	STS		
INT	STI	LAT		ALI	SED		
COM	OPE	CAT		ARY	ING		
OMM	ELE	ECT		ERS	RES		
MIN	SIT	ILL		IST	STI		
PRE	EDU	EST		ING	SMS		
ONS	ANI	TAT		ATE	TIE		
NVE	OMM	ESS		ALS	ZAT		
DIS	PLI	SIT		ABL	ZED		
TER	ROD	VIS		NAI	SIN		

How to do it

To generate patterns first select the item you wish to study in the main chargram display 414, then in the menu, choose *Compute* | *Patterns*.

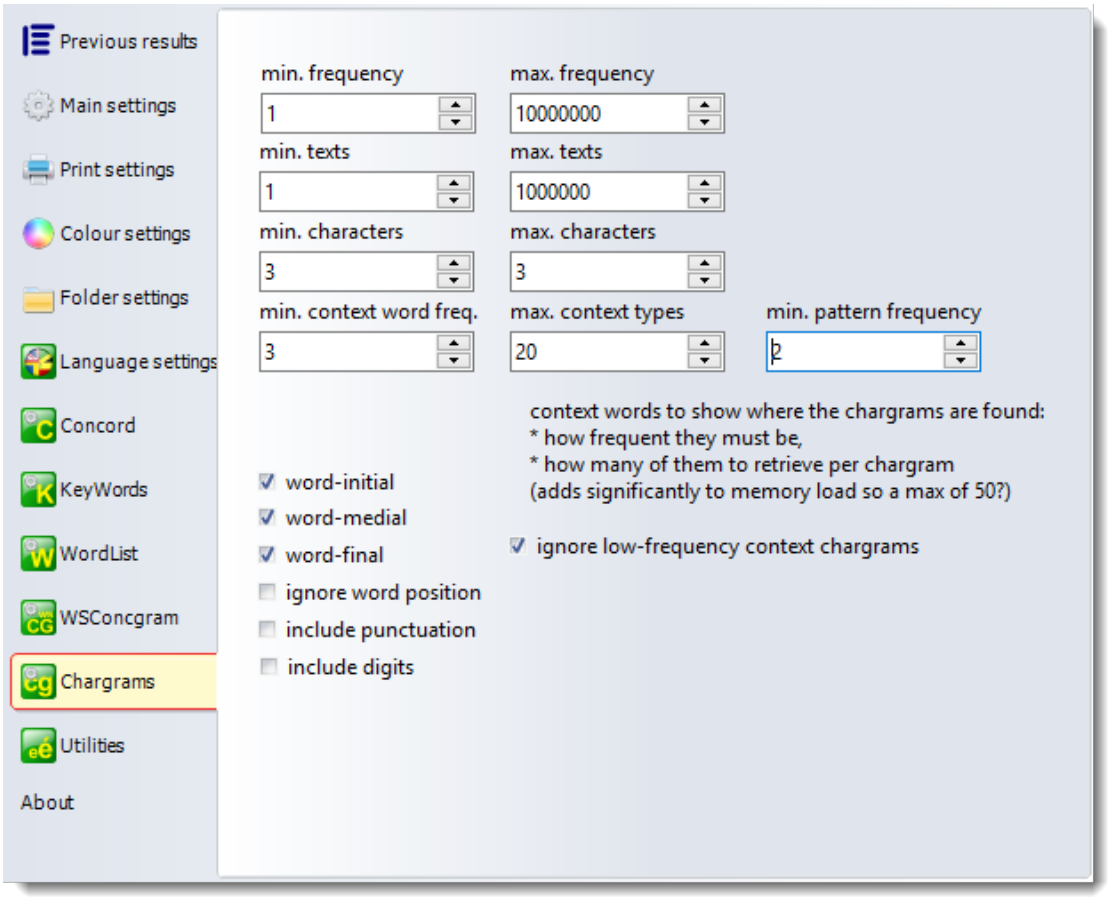
The procedure uses the same settings 418 as in generating chargrams, with a minimum frequency for the list. You can see the frequency of any item by selecting it.



See also: Computing chargrams 411.

10.3.5 settings

Settings are found in the main Controller.



You can set minimum and maximum token frequencies for the chargrams to be included in the results, a minimum and maximum number of texts they must appear in, the length in characters (e.g. 3 to 4 characters).

Context words

As the chargrams are selected, note is taken of the word which they are found in. Here you can determine a minimum number of times for each chargram to appear in a given word for that word to be listed, and a maximum number of context words per chargram to be collected. (Storing lots of extra words will use up system memory so a default of 20 or 50 may be reasonable.

Word position

By default chargrams in all three positions (word-initial, word-medial and word-final) will be collected. If you check the *ignore word position* box, word positions get merged.

Include punctuation

Allows chargrams of all characters (symbols, punctuation etc.) to be included. Spaces get replaced by underscores.

Include digits

Allows chargrams of digits as well as alphabetic characters.

Ignore low-frequency context chargrams

This setting allows us to filter out any chargrams which do not occur in many contexts. As shown here, any chargrams not occurring in at least 3 context word types will get eliminated. If, for example, a chargram has been found in 5 context word-types then the chargram is included in your list. But if in only 1 of these it is found occurring at least 3 times (i.e. found in the same context word-type recurring in the texts at least 3 times), you will see one context word only in the *contexts* column

min. pattern frequency

This setting lets you restrict low-frequency patterns 416.

See also: chargrams display 414

10.4 Corpus Checker

10.4.1 Aim

The purpose is to check your corpus for corruption, relevance and duplicates.

Examples:

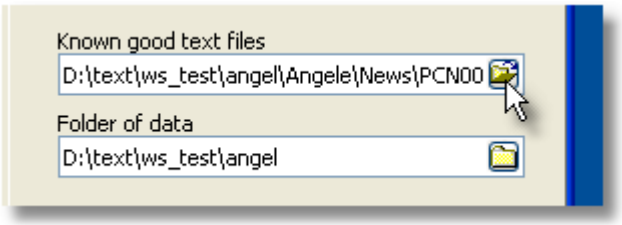
- it has got corrupted so what used to be good text is now just random characters or has got cut much shorter because of disk problems
- it isn't even in the same language as the rest of the corpus
- it is a copy 425 of another text file or has very similar text 422
- it contains too much boilerplate text 426
- it is or is not relevant 429 to a particular area of enquiry

The tool works in any language. It checks corruption by using a known sample of good text (in whatever language) and comparing that good text with all your corpus.

See also : How to check corruption 420

10.4.2 checking corruption

1. Choose a set of "known good text files" which you're sure of. The program uses these to evaluate the others.



When you click the button for known good text files, you can choose a number. You might choose 20 good ones so as to get a lot of information about what your corpus is like.

2. Choose your corpus head folder and check the "include sub-folders" box if your corpus spreads over that folder and sub-folders.

3. The program will anyway look out for oddities such as a text file which has holes 574 in it, eg. where the system thinks it's 1000 characters long but there are only 700.

4. If you check the "digraph check" box it will additionally check that the pairs of letters (digraphs) are of roughly the right frequency in each text file. For example there should be a lot of TH combinations if your text is in English, and no QF combinations. If you are working with a corpus in Portuguese and your text files are in Portuguese too, of course the digraphs will be different, and TH won't be frequent. The program ignores punctuation.

5. If you are doing a digraph check you can vary certain parameters such as how much variation there may be between the frequencies of the digraphs (a sensible setting for "frequency variation per 1000" could be 30 (in other words 3%)), and "percent fail allowed" (which might be set at say 25 -- this means that up to 25% of the digraph pairs may be out of balance before an alert is sounded).

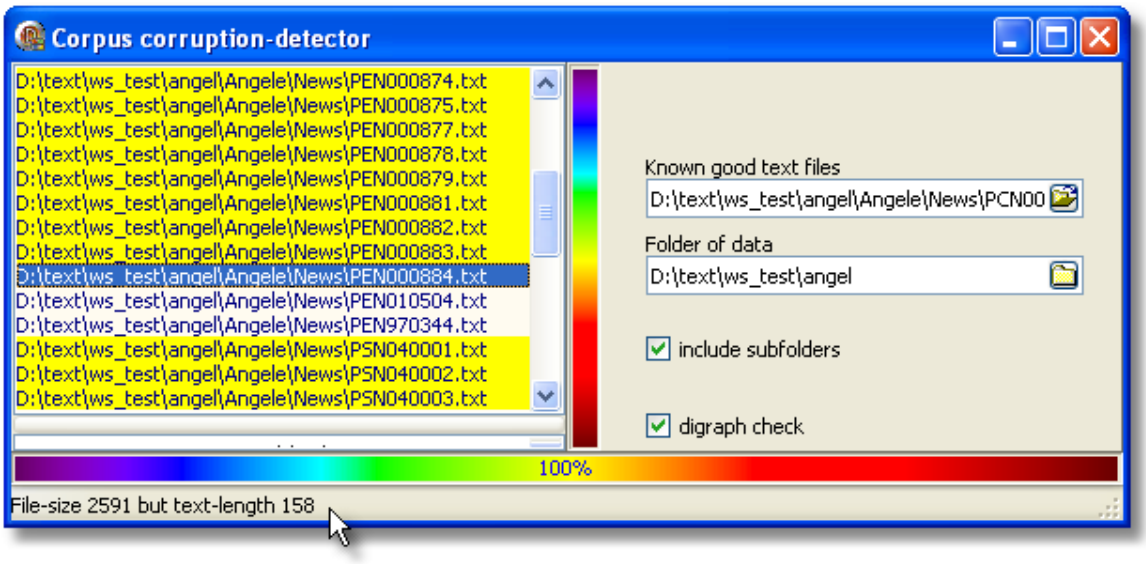
6. Press Start.

You will see the progress bar moving forward.

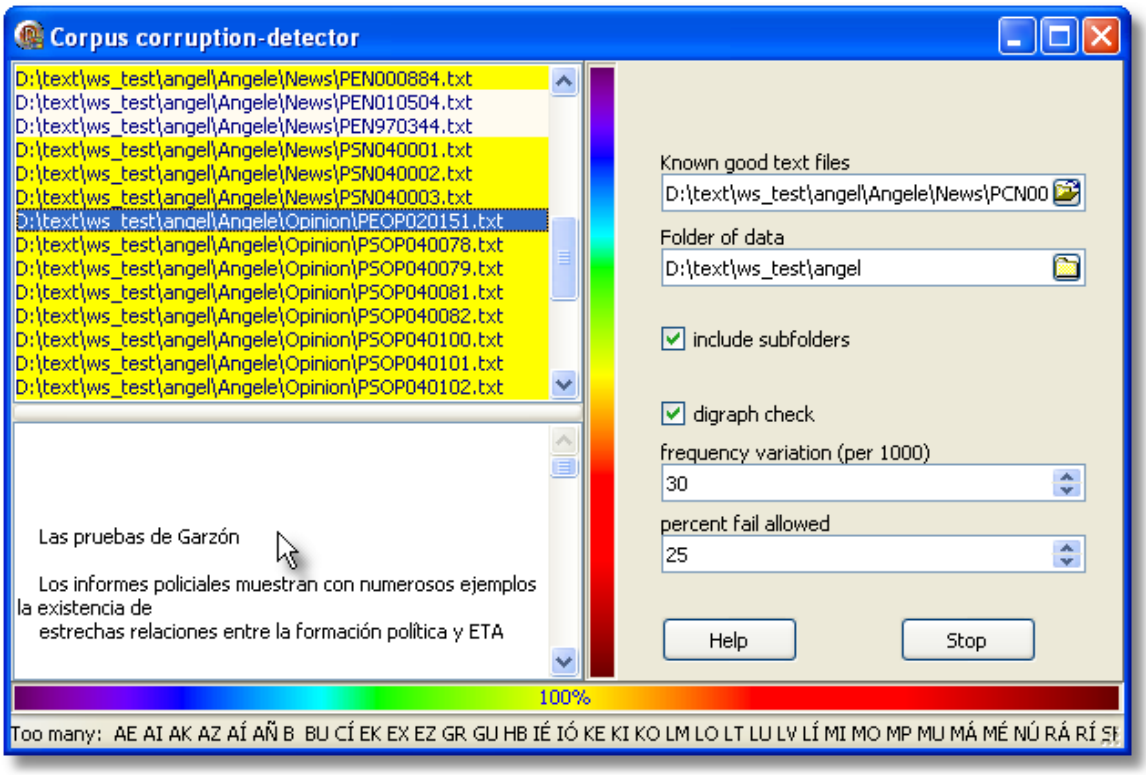
If you see a file-name in the top-left box, a click on it will indicate why it was found questionable. Double-clicking it will open up the text in the window below so you can examine it carefully.

Filenames of possibly corrupted texts are yellow if the basic check fails, and cream-coloured if the reason is because of a diagraph mis-match.

In the screenshot, PEN000884.txt is problematic because the file-size on disk is 2591 (there should be 2591 characters) but there are only 158, as shown in the statusbar at the bottom.



In the case of PEOP020151.txt, the text appears below (after double-clicking the list),



and the status bar says the tool has found an imbalance in the digraphs. The text itself has a lot of blank space at the top but otherwise looks OK (it is supposed to be in Spanish) but the detector has flagged it up as possibly defective.

10.4.3 find duplicates

We offer two utilities for finding duplicate text files:

- o same name but in different folders 425
- o any name examining the contents 422.

10.4.3.1 similar content

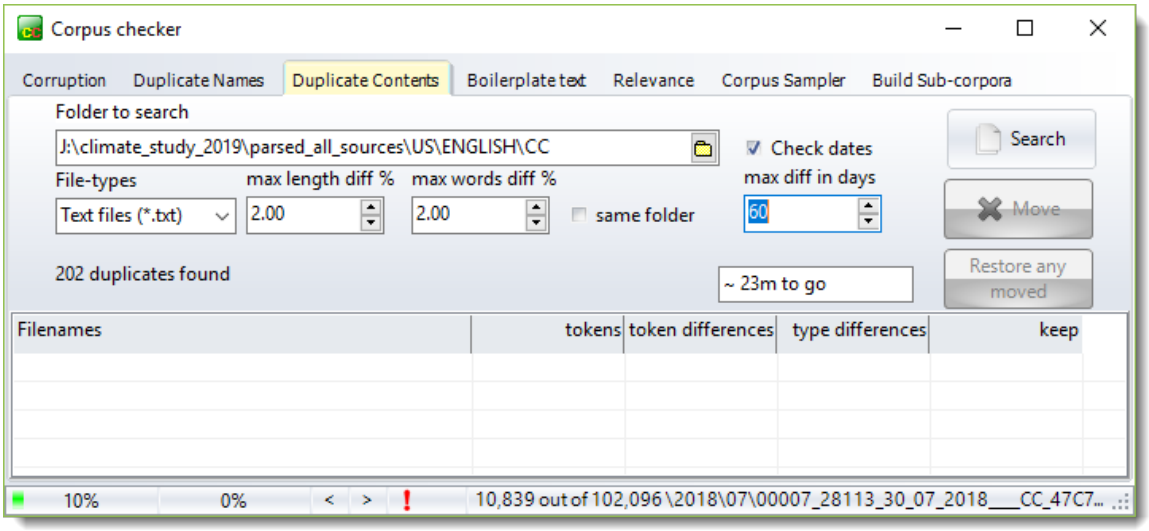
The point of it

News texts are often edited slightly and re-published a few days later or in a sister publication.

The idea is to be able to check whether you have virtually duplicate text files, i.e. with the same or similar contents. This search compares all text files with each other regardless of their file-names. If duplicates are found you can move and rename them so they can be avoided in further research.

How to do it

Specify your *Folder to search* and file-type(s). Check the difference settings, then simply press "Search". Search will go through that folder and any sub-folders and will report any duplicates found. Note: it will take time, as it needs to compare each text file with all the others in those folders. The comparison checks the vocabulary used in each pair of texts.



Here the status bar shows there were over 100,000 text files, 10% have been checked and 202 near-duplicates found. The remaining time is estimated at 23 minutes.

Settings

An intrinsically slow process can be speeded up by reducing the number of text files which get compared with each other.

Max length diff = how much difference in length will you accept in a comparison? The bigger any length differences the slower the process. In the case above a setting of 2.0% was used, so any a text of exactly 100Kb in size would only get compared with others between 98 and 102Kb in size.

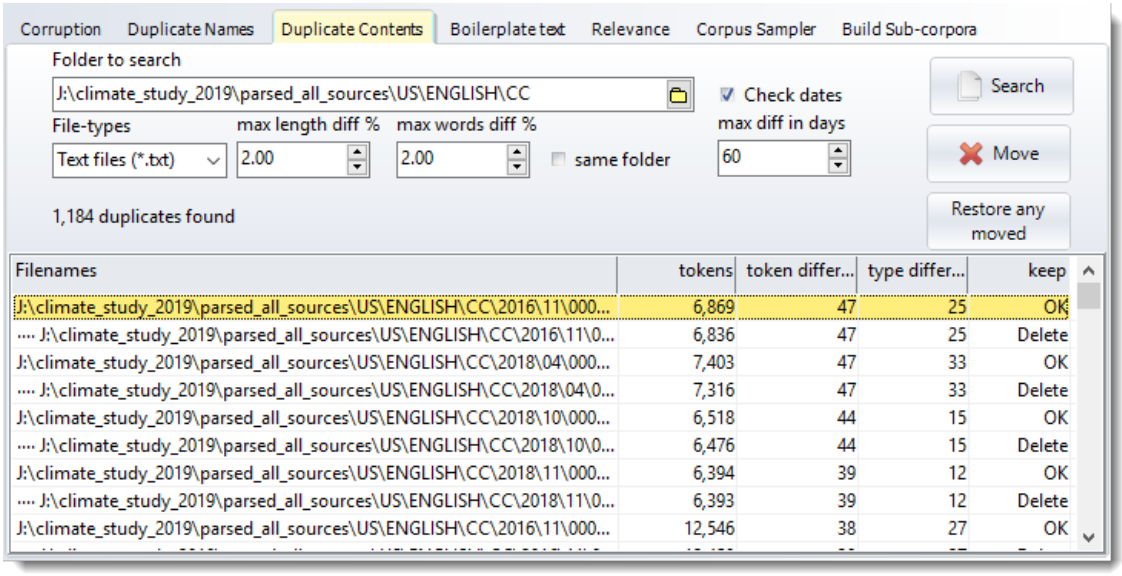
Max words diff = how much difference in types or tokens will you allow and still consider the texts duplicates?

Check dates: if this is selected, you can decide how many days apart text files can be.

Same folder: if this is checked, only text files in the same folder get compared. In the study above that would have meant only texts of the same month got compared.

Pause the process? Well, you can stop it (press the red exclamation mark). To re-start at a specific point, hold down *Ctrl* as you press the *Search* button and you can re-start at 25% or wherever you wish.

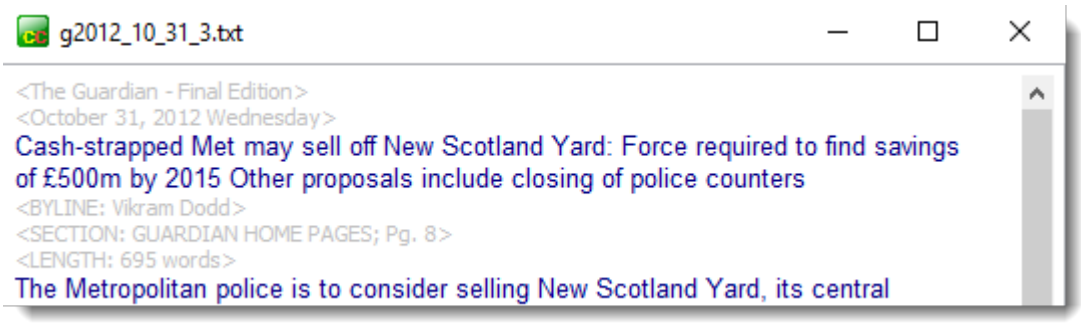
Results

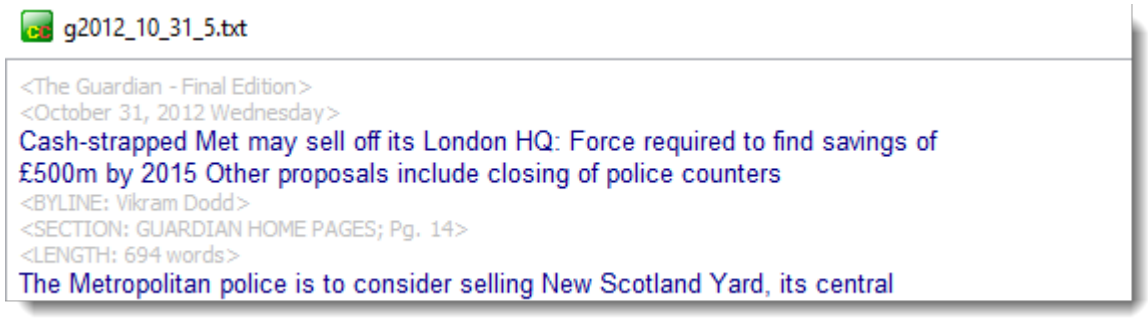


The highlighted line shows two text files with less than 2% difference in size and 60 days apart. The longer was 6,869 running words in length (discounting mark-up). It's 47 words longer and has 25 different types. By default, the longer one is marked for keeping in the *keep* column, where the shorter is marked for deletion (see *Remove* below).

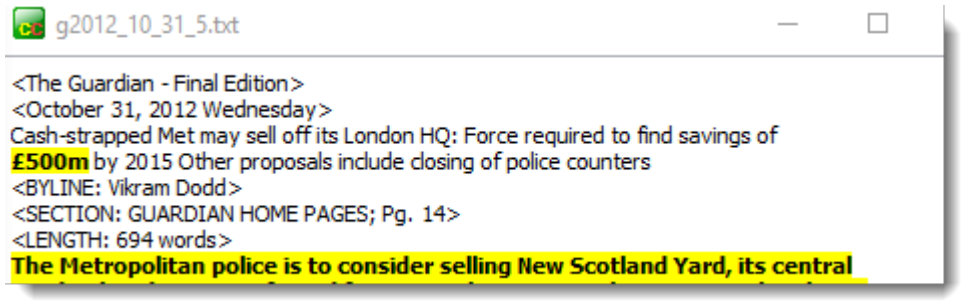
The most different sets are listed at the top. In this case, the top text had the same basic story in the same newspaper (New York Times) but one day later, and gives more information on the contributing journalists and a few links to related articles.

Right-clicking and choosing *Show* reveals minor differences in any two or more texts:





Right-clicking the latter window and choosing *Highlight similarities* gave this:

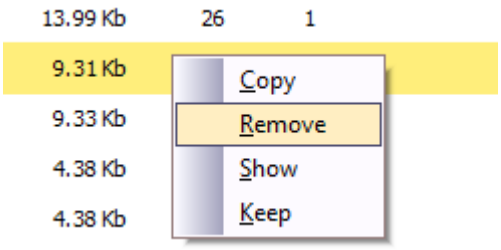


Angle-bracketed lines are not compared. In the rest of the text, identical lines get coloured yellow. Most of the rest of the text is coloured yellow. Here what has happened is that an editor has changed **`New Scotland Yard`** to **`its London HQ`**. It looks as if minor edits were performed but otherwise text 3 and 5 are duplicates.

Copy, Move, Show, Keep

When the search finishes you can press *Move* to filter duplicates to a *moved* sub-folder. Use the *Restore any moved* button to go through all the set, putting any moved back where they belong. (You may have to close the Corpus Checker first.)

Right-clicking the display of duplicate files gives these options:



To view, choose *Show* in the right-click menu. Or double-click to open any file in Notepad. *Copy* produces a list of the results which you can paste elsewhere. *Remove* simply moves all the duplicates which are marked for deletion, to a sub-folder of where they were before, called "moved", and changes the file extension to **`.moved`**. To change which duplicate is kept, select the one you do want and press *Keep*.

How does it check duplication?

The process goes through all text files in the folder chosen plus any sub-folders of it, and compares each one with all the remaining text files. If differences in both files' sizes are within the *max length diff* (and optional date or folder settings), the two text files may be compared. A rough check is first done, to see whether a chunk of a certain number of characters is found in each of the two text files under consideration. If so, word lists of each are computed, ignoring sections marked in angle-brackets. The frequency of each type is compared, giving any difference in tokens and types. If the differences in token frequencies or type frequencies are within the *max words diff* percentage, the pair are identified as near duplicates.

Any text will only be listed as a duplicate of one other text.

This procedure is fairly thorough but will take time because so many text files get compared. A time estimate is shown. (Note: it processes the longest text files first and should speed up as it gets to the shorter ones.) It assumes that the word list of a text which matches the word list of another text is a good sign of content duplication, regardless of word order.
The date comparison uses file dates 440 (which you can set appropriately for text files).

To highlight similarities, the procedure is to take each line of text in one of a pair (unless it's angle-bracketed) and see whether it is also found in the other. If so, it'll get coloured yellow.

See also: duplicate file-names 425, count files 450.

10.4.3.2 same names

The point of it

The idea is to be able to check whether you have files with the same name in different folders. You may often make copies of files and a few weeks later cannot remember where they were.

By default this program only checks whether the files it is comparing have the same name but dates and file-size can be compared too. It handles lots of folders, the point being to locate unnecessarily duplicated files or confusing reuse of the same filenames.

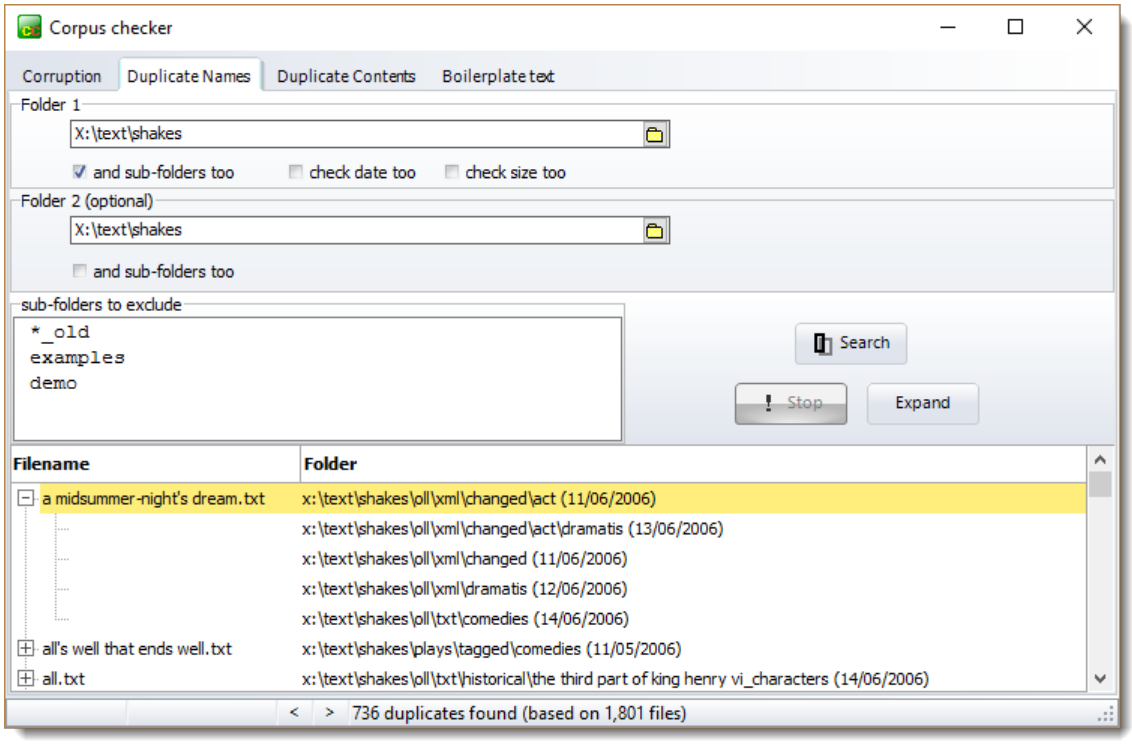
How to do it

Specify your Folder 1 and simply press "Search". The *Duplicate Filenames* procedure will go through that folder and any sub-folders and will report any duplicates found.

Or you can specify 2 different folders (e.g. on different drives) and the process compares one set with the other.

Sub-folders to exclude

Useful if there are some sub-folders you know you're not interested in. In the example below, any folder whose name ends **_old** or whose name is **demo** or **examples** will be ignored as will any sub-folder below it.



In the window below, you will find all the duplicates listed with the folder and date. In the example we can see there are five versions of **`a midsummer-night's dream.txt`** in different shakespeare folders. In brackets is shown each file's date.

See also : duplicate contents 422

10.4.4 boilerplate text

The point of it

The aim here is to find repeated chunks, such as can get caused

- if someone has inserted a paragraph twice by mistake
- by plagiarism
- by re-writing and editing text
- in copying and pasting.

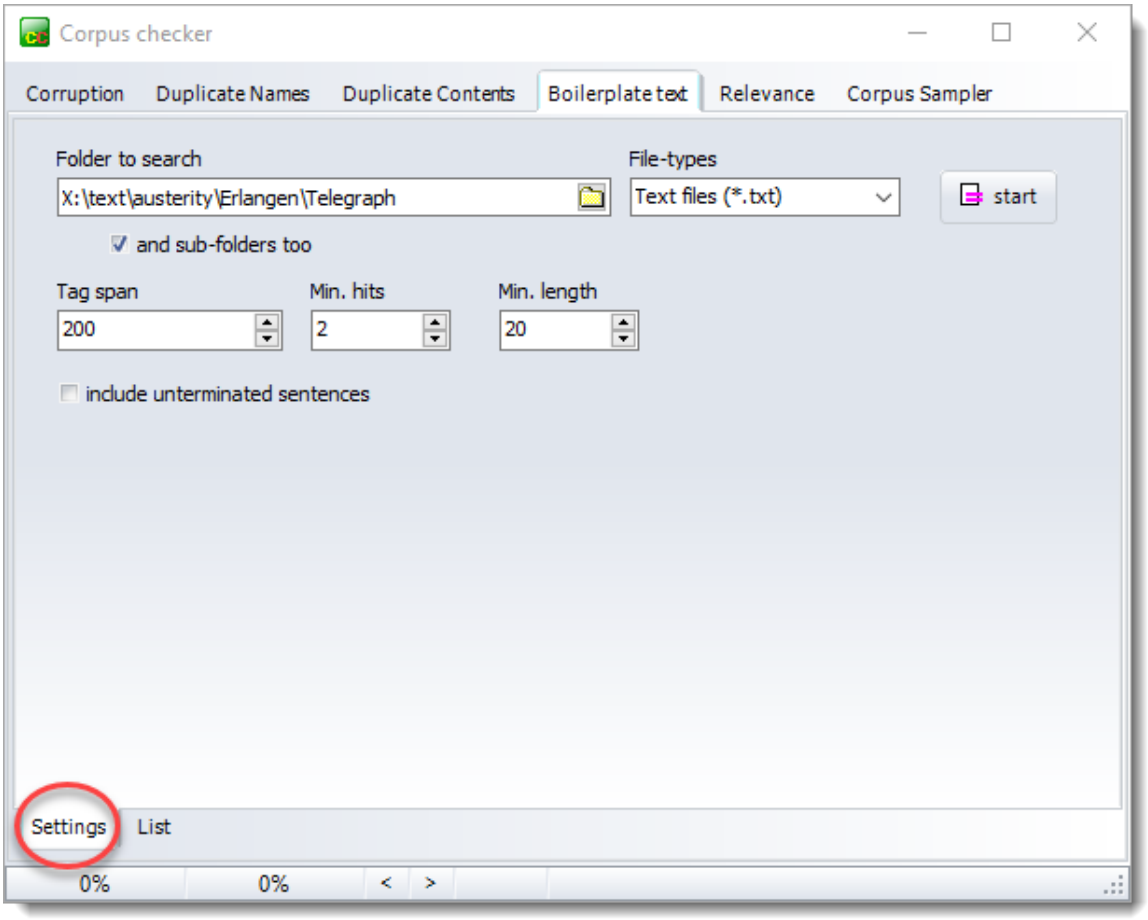
The procedure looks essentially for repeated sentences and headings in a whole lot of texts.

How to do it

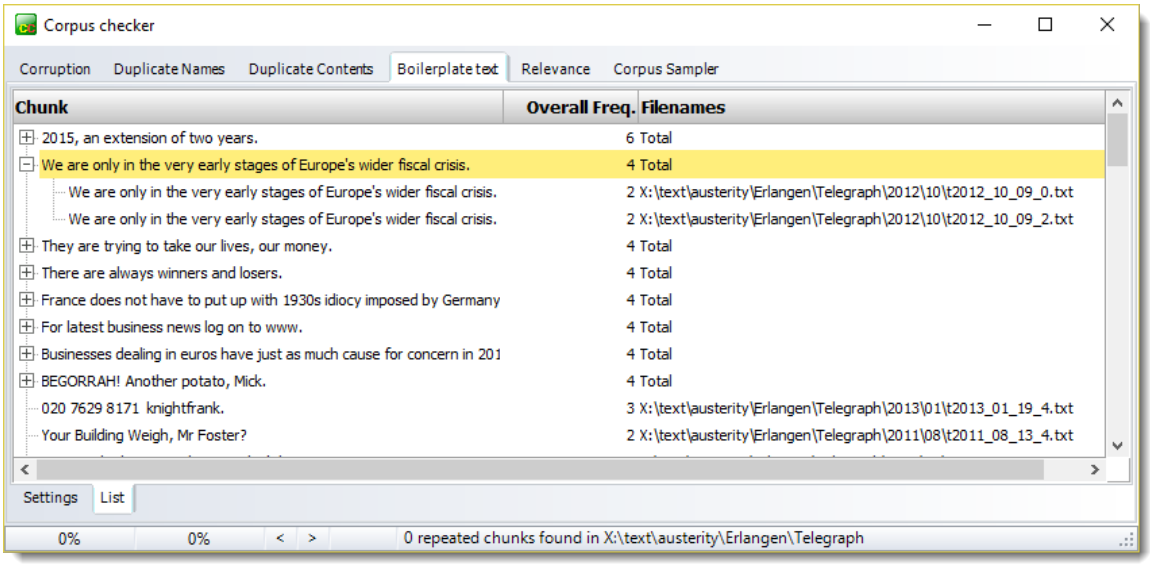
In the Settings tab with the yellow oval below, choose a folder. It will be searched as will all its sub-folders, except any called *filtered*.

Choose the *file-types* to search (default ***.***) and a *tag span* 159 such as 200 characters, since mark-up gets ignored in this search. Set the minimum number of *hits:* the number of repetitions which you're interested in seeing in any text file. Min. length is the length of any repeated chunk. *Include unterminated sentences*: includes headings.

Press start.

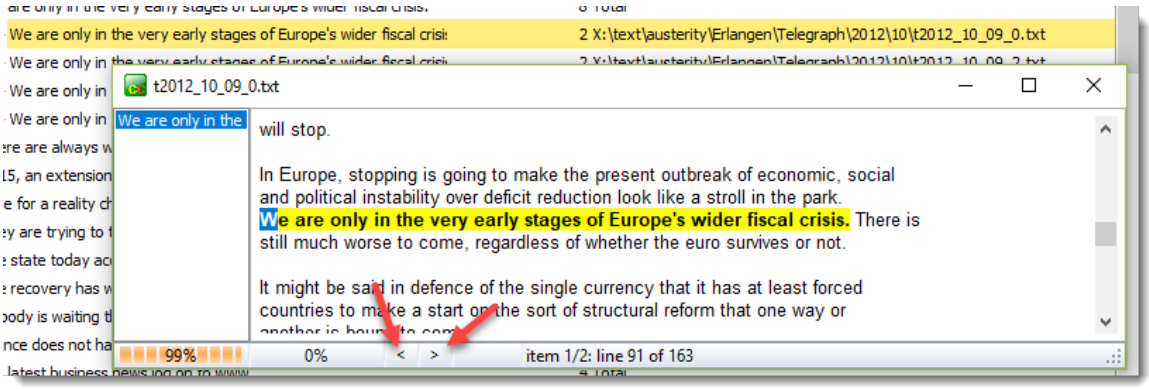


You may get results like this:

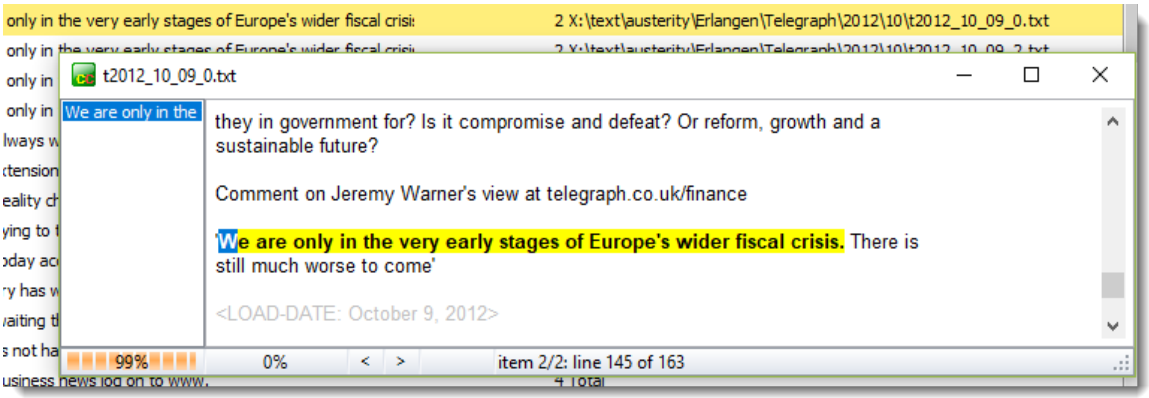


In the first case a chunk has been found repeated in 2 different text files both of the same date.

Right-click to see the text in question or to copy the list to the clipboard or to Excel.



Press either of the two buttons shown by the red arrow to jump from one highlighted chunk to the next.



The top context is in the middle of the text and the sentence gets repeated at the end, possibly as a caption to a picture.

See also: duplicate file contents 422, corruption check 420, duplicate file-names 425,

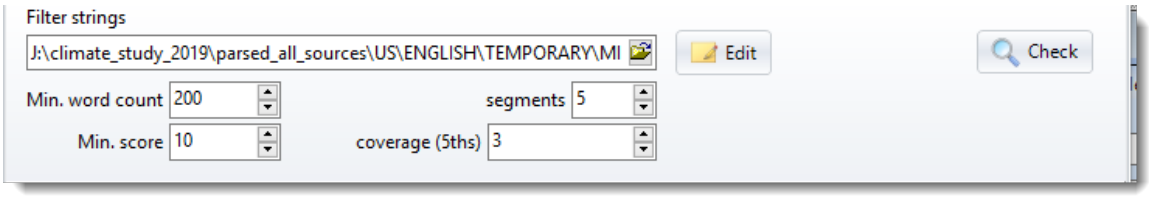
10.4.5 relevance check

The point of it...

The aim is to filter your corpus by checking which of certain words or phrases are found or not found in each text. It operates a scoring system. You specify words or phrases which you believe are typical of the field you're investigating and can specify some which you see as unwanted distractors. Text files which score highly can then be copied or moved to a location of your choice.

Settings

Choose a list of filter strings, a minimum word count and a preferred minimum score.



coverage: you can choose a number of equal-size segments of each text which any hits must be found in.Here, the settings mean the texts will be segmented into 5 sections (the first 20%, the second 20% and so on) and any three of them must have at least one hit.

Filter strings syntax

Any filter string which starts with ~ count negatively.

As each is found it will score 1 point but you can increase that by adding a value so in these filter strings:

```
cutting spending
spending cuts=2
spending review
~carbon
~immigration
```

here, **spending cuts** counts twice as much as the others.

More complex searches

In addition, you can require context words. For example CHANGE with a left context requirement of CLIMATE.

The syntax for that is

main word<tab>left context word<tab>right context word<tab>context span (characters)<tab>value if found

e.g.

```
change<tab>climate<tab><tab>40<tab>6
```

which means search for **change** with **climate** within 40 characters to its left, and give it 6 points if found.

Finally, you can require a number of other search words to be found as a condition. For example in a search on climate change, terms like **oil, gas, electricity** might be relevant. In this case the number 4 would mean, *only include any hits of* **oil** *if at least 4 of my other terms in the list are found, per 1,000 words of text*.

```
oil<tab><tab><tab>4<tab>20
```

which means search for **oil**, no context requirements or context span, but only include it if 4 other search terms are found in that text (proportionally more if greater than 1,000 words). If it is found give it 20 points.

An example:

; word	left context	right context	context span	others required	value if found
crisis	climate		50		20
fund	climate		50		20
warming				5	20
global				5	20

Tip: enter your terms in a spreadsheet, then export to plain text separated by tabs.

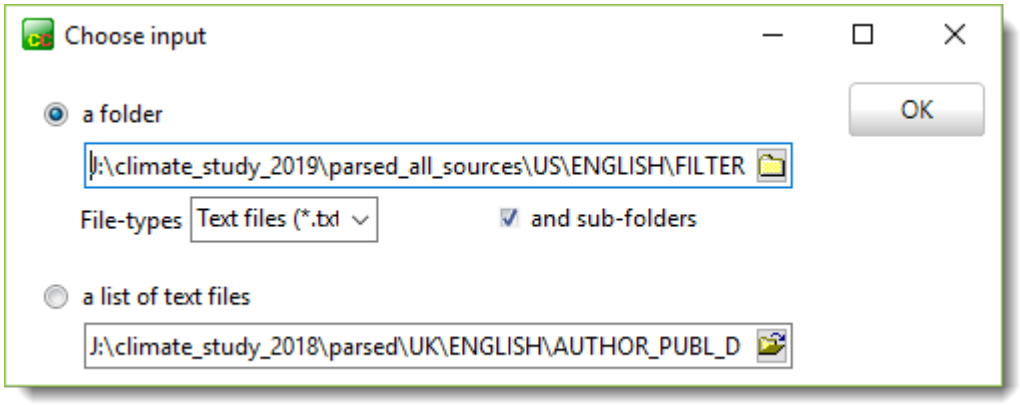
Headings and terms you wish to ignore?

Put ; or // at the left of the line. Examples:

```
// my comment to myself
; a reminder
```

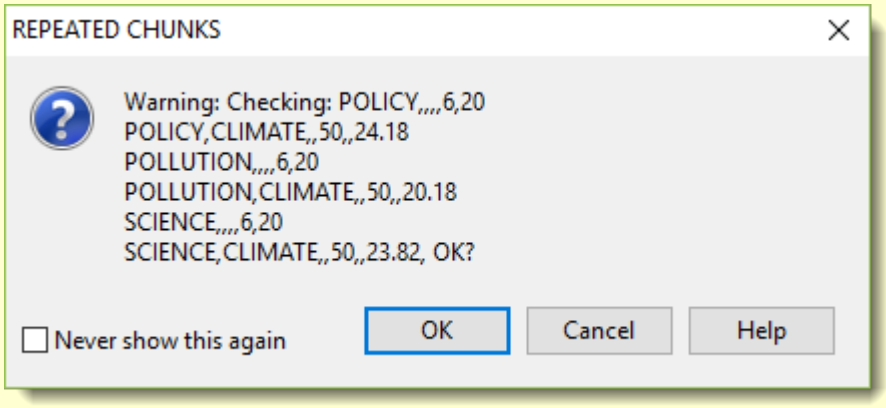
Which texts to consider?

When you press the button you will get a choice between all the text files in a folder and sub-folders, or a previously-made list, such as that created in the Corpus Sampler 434 procedure.



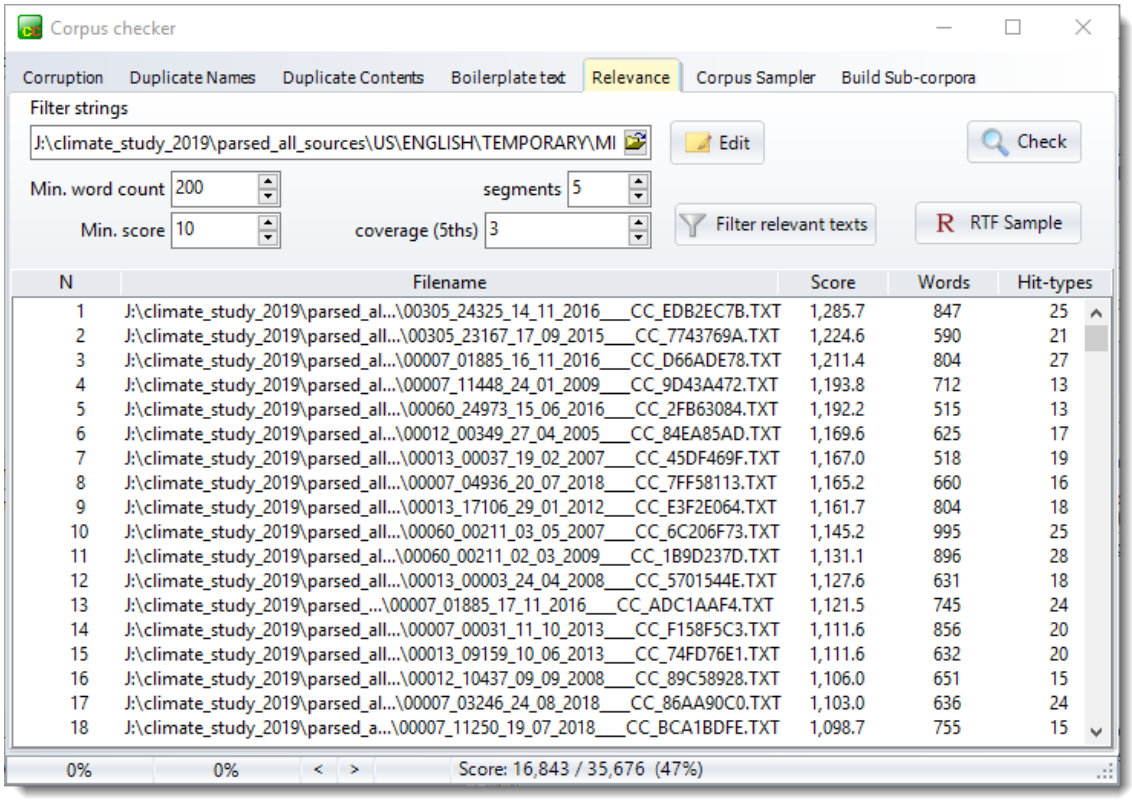
repeated chunks

A warning like this pops up if you repeat filter terms.



Here we have a filter which seeks **POLICY** and also seeks **POLICY** with **CLIMATE**.

Display



The display shows scores, number of words and number of different hit-types found. In the status bar at the bottom you see that this search filtered out just over half of the 35,676 text files. The score is just a number. It will vary according to how much value you attach to a search-term, the number of search terms you seek. In the search above the values averaged about 20; there were 91 search terms.

Example

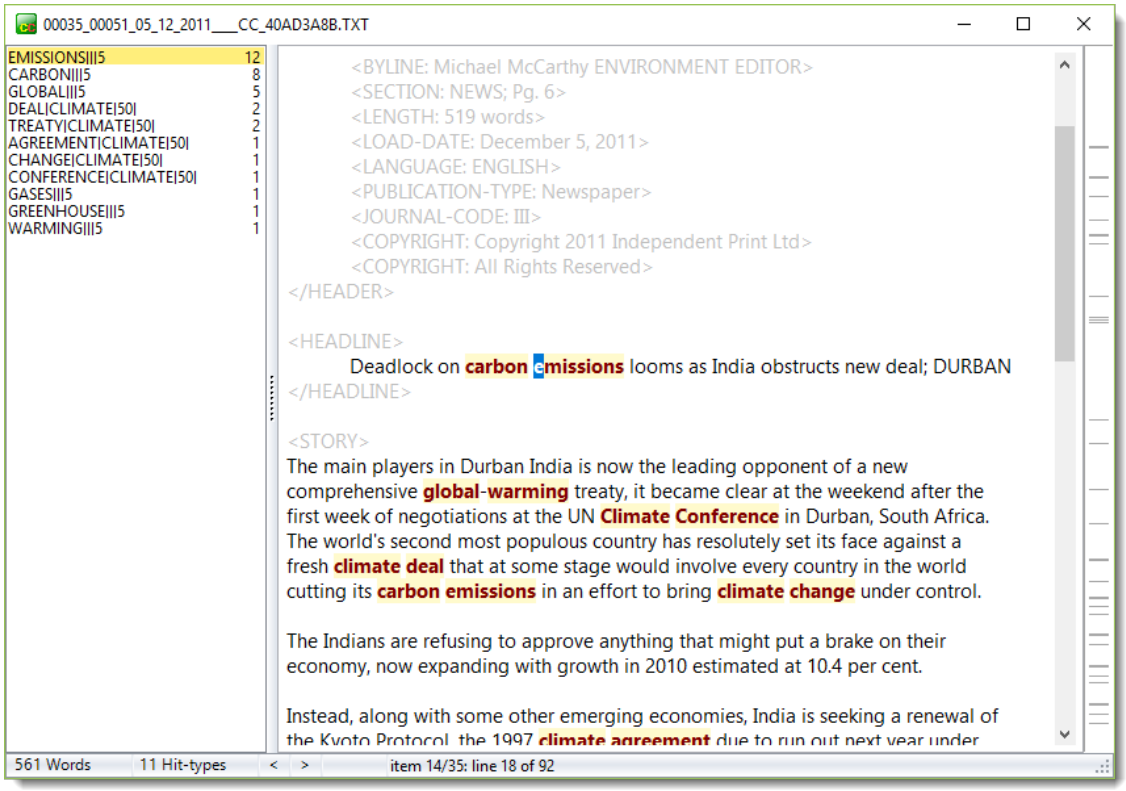
I was studying austerity in news text. Lots of articles mentioned austerity, sometimes incidentally. And I wanted to study austerity in Britain but a lot of articles concerned Greece. So my filters had terms like **`cost-cutting, UK`** etc. and my negative filters included **`Greek, Greece`** etc. To get a suitable corpus I wanted quite a lot of the positive terms I preferred and few of the negative ones. After the relevance check was done I was able to filter out most of the texts leaving only ones which were much more relevant to my enquiry.

Filter relevant texts 434 button

RTF Sample 436 button

See the text

To see one of the texts simply select it and double-click (or right-click and choose *Show this text)*. When it appears you will again be able to right-click in order to save it as RTF, grey out any < > sections, highlight the positive filters, etc. After highlighting all, we get



You see the search terms coloured, and to the right a dispersion plot showing where they appear in the text.

Limitation

The procedure uses search-terms. It doesn't actually understand the text. All it can do is give a higher score to the presence of positive terms and reduce the score if negative ones are found. Texts about the environment don't necessarily contain the word **`environment`**!

Choosing your relevance filters

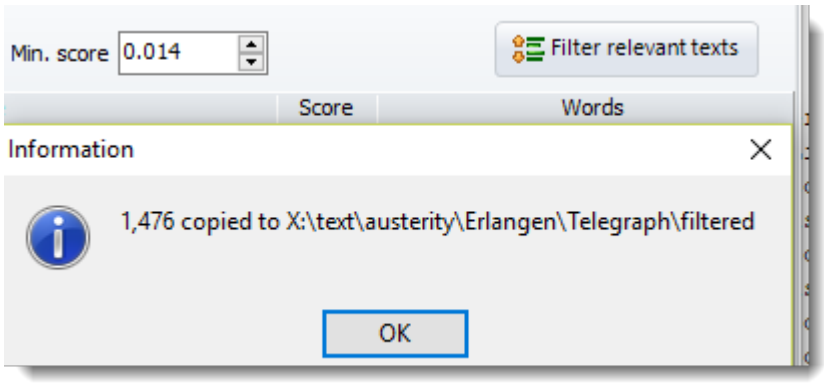
A useful idea is to compute the key words and key clusters of your imperfect corpus first. That will help you find words and phrases that characterise your corpus. Use some of them plus any others you think will be plausible. Also, read a sample 436 of texts carefully to check what sort of corpus you really got.

Finally try the relevance filter. You can sample the texts to see how well you're doing. Edit the relevance filters to refine them.

See also: RTF sample 436, corpus sampler 434, which helps you separate the desired sample out.

10.4.5.1 Filter relevant texts

The ones visible scored above the *Min. score* setting. Press *Filter relevant texts*, and those scoring highly enough get copied to a sub-folder called "filtered".



10.4.6 corpus sampler

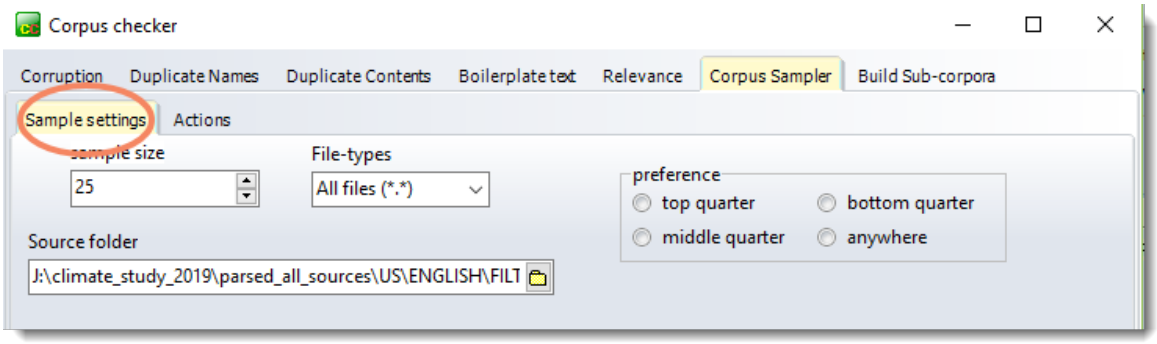
The point of it

Texts downloaded from the Internet often seem to promise one thing but in fact contain another... A news story may refer incidentally to an issue you are interested in, but really be mainly about something else.

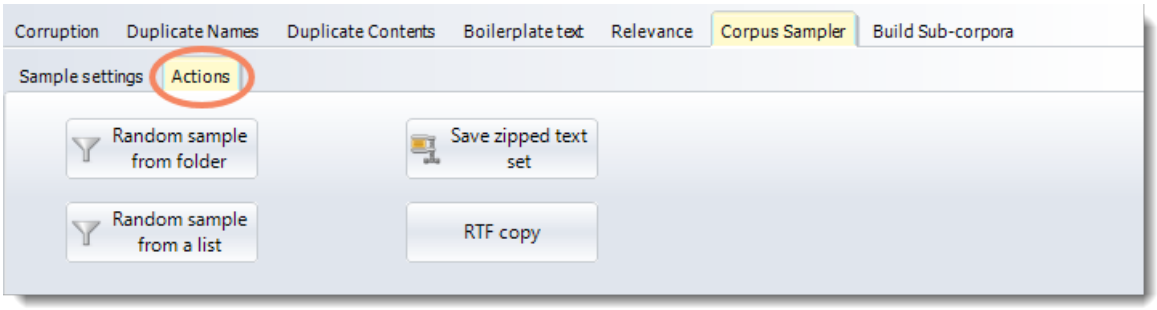
Use the sampler when you need to read the individual texts in your corpus to check whether they meet your purposes. This utility finds a random set and can save that for sharing with others so you can agree on criteria.

How to do it

In the *Sample settings* tab, choose the sample size desired, the types of text file and the folder they are in. Decide whether you want a sample from the top, bottom or middle of the collection (as sorted by file-date if from a folder, or from a list of files you already have).



Now in the Actions tab,



press *Random sample from folder* if you want the program to go through a set of texts in a folder (and its sub-folders), or *Random Sample from a list* if you have a plain text list. The procedure looks at all the text files, and picks out that number at random. (If there are aren't enough you'll get what there are.)

Here the random procedure has chosen *anywhere* within the total.

N	Filename
1	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2005\02\00008_00210_15_02_2005___CC_5EB36416.TXT
2	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2006\07\00120_00003_23_07_2006___CC_47E546B8.TXT
3	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2006\10\00007_00147_31_10_2006___CC_49904428.TXT
4	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2007\03\00060_05079_23_03_2007___CC_F2F9ECB3.TXT
5	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2007\03\00118_04884_08_03_2007___CC_96B59620.TXT
6	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2007\05\00118_00003_24_05_2007___CC_48934A4F.TXT
7	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2007\07\00007_00003_07_07_2007_21_CC_17C4137C.TXT
8	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2007\07\00013_05850_15_07_2007___CC_77DE7D5F.TXT
9	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2007\11\00013_07754_27_11_2007___CC_ADBAB0B2.TXT
10	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2008\07\00120_10143_24_07_2008___CC_8B018E3D.TXT
11	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2008\11\00060_00583_07_11_2008___CC_CF73D03A.TXT
12	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2008\11\00060_04216_17_11_2008___CC_8C4F8333.TXT
13	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2011\07\00120_15329_14_07_2011___CC_E5BFE845.TXT
14	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2012\01\00007_09735_10_01_2012___CC_594F546A.TXT
15	J:\climate_study_2019\parsed_all_sources\US\ENGLISH\FILTERS\pubs_OK\CC\2012\07\00007_01681_22_07_2012___CC_2EED2585.TXT

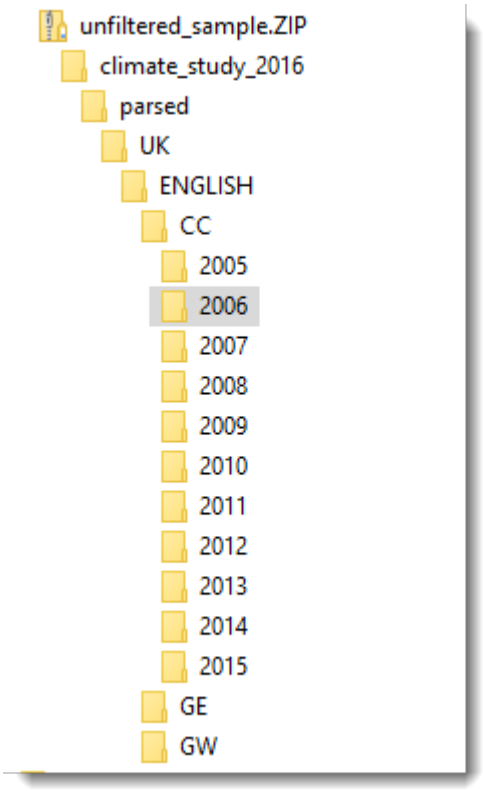
I have highlighted the text file dates in yellow: they are still in date order but only 25 were selected from a set starting in 2005 and ending in 2019.

RTF Copy 436 button

Lets you save some or all in RTF format, so that you can read and study the texts in Microsoft Word or similar.

Zipped text button

This generates a zipped file, with all the original folder structure of the sample texts preserved:



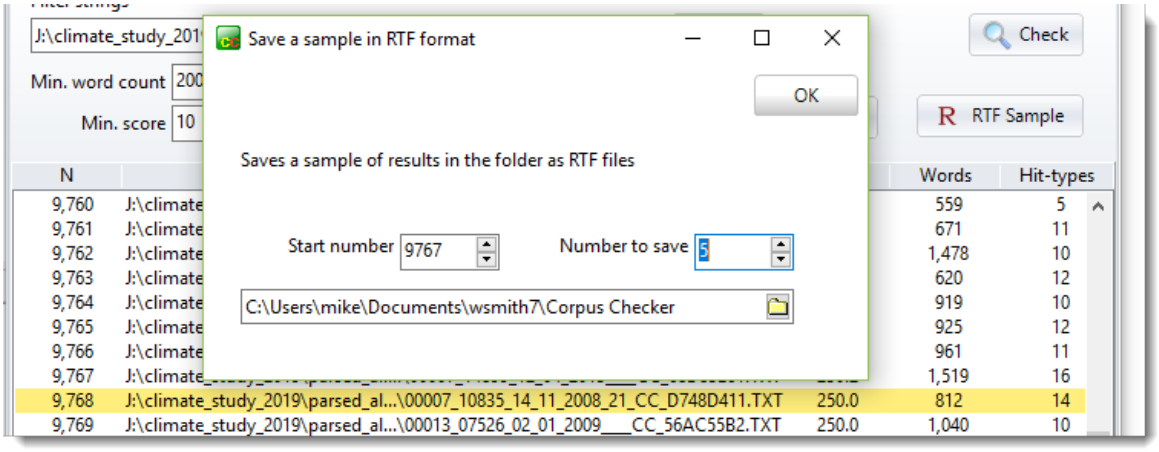
In the .zip you find a separate copy of each text file, as well as a large text containing a copy of each text glued one after the other. It also contains a plain text list of the text files ready for you or fellow researchers to note down observations about each text. You might find it useful to copy the latter into Excel in your record-keeping.

See also: relevance check 429. And File Utilities 10 has a tool for moving or copying 448 the structure of text files.

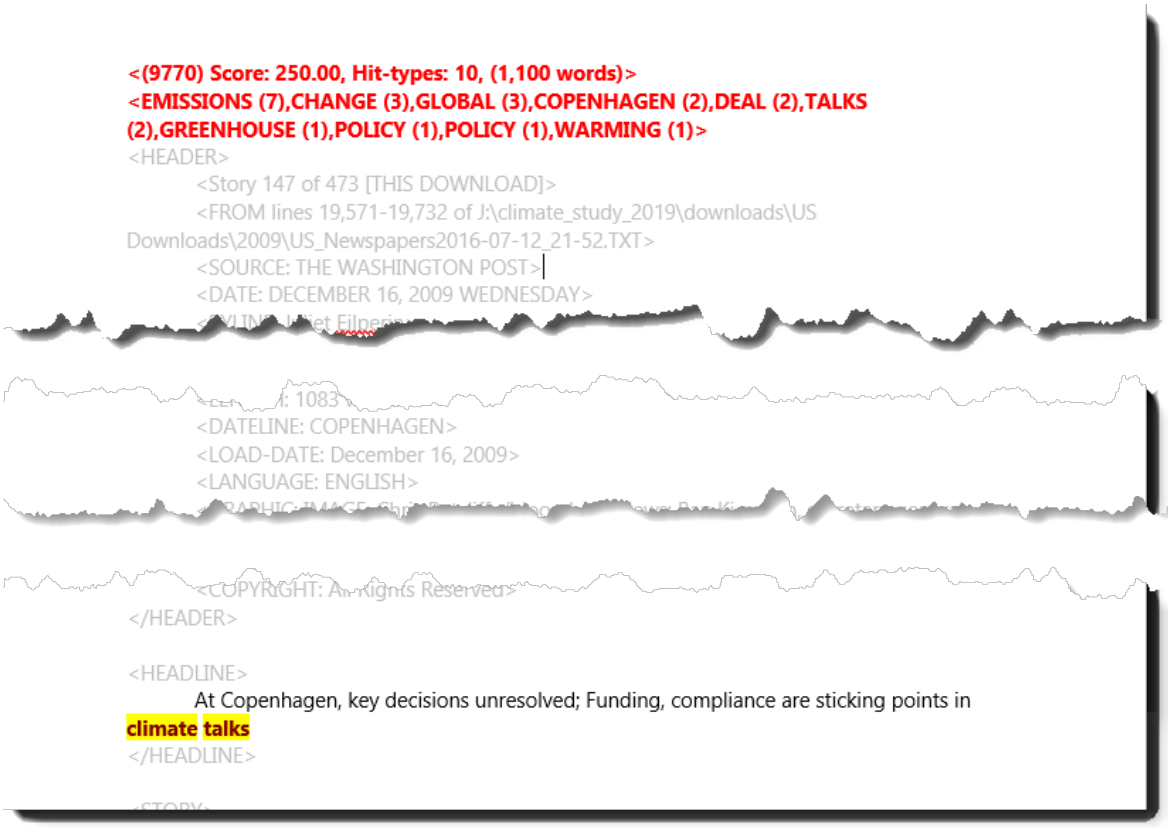
10.4.7 RTF copy

The idea is to save a sample of texts in RTF format with colours to show the source text and how many filter terms it has.

Here I chose the scores of 250.0 and saved a sample of 5 texts.



The results look like this:



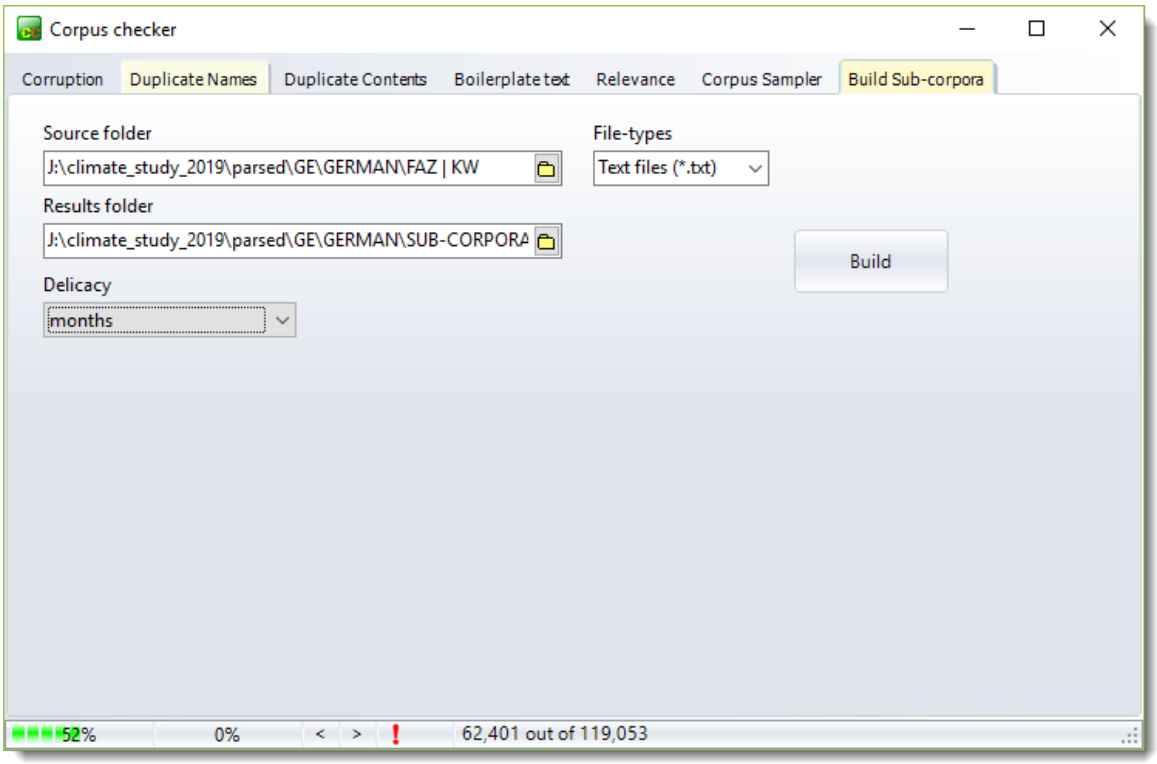
You get relevant score details in red at the top, the headers greyed out, then the text with all filter terms highlighted. It looks as if this is a climate change text.

See also: See the text 432

10.4.8 build sub-corpora

The point of it

It can be convenient to copy all the text files of one given month into one single text. It uses the date 149 of each text. This function reduced 120 thousand single articles between 2005 and 2018 into 168 monthly texts, easier to handle when seeking trends.



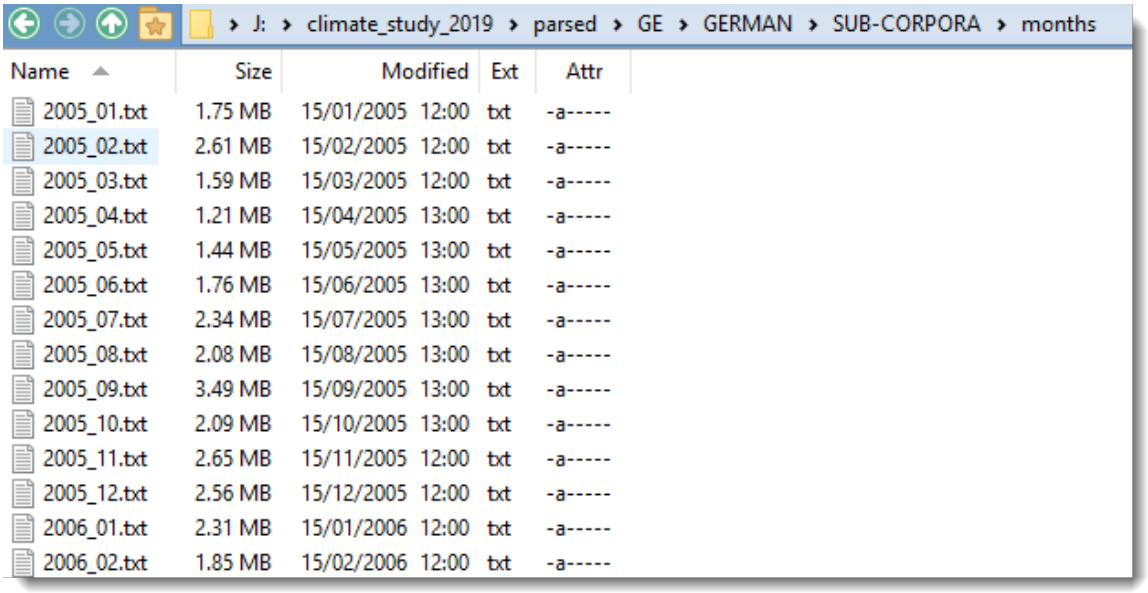
How to do it

Choose a source folder, the file-types and a suitable results folder, and choose the delicacy required: years, months, weeks, days are the obvious choices.

In the screen-shot you can see I added **|** and a further folder-name (**KW**) after choosing a source folder **j:\climate_study_2019\parsed\GE\GERMAN\FAZ.** This allows the program to search within both the FAZ and KW folders, both of which stem from **j:\climate_study_2019\parsed\GE\GERMAN**.

Results

As each text file is found it gets added to the collection for the appropriate year/month/week/day in a folder of that name. This is as partial view of what I got:



10.5 File Utilities

10.5.1 index

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

Compare Two Files 440
Dates and Times 440
Move or Copy files 448
Delete 450
Rename 451
Lemma file building 444
Tags 446
Highlighter 442
Splitter 453
Joiner 457
File Chunker 458
Find Holes 459: for "holes 574" in text files

10.5.2 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 458 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

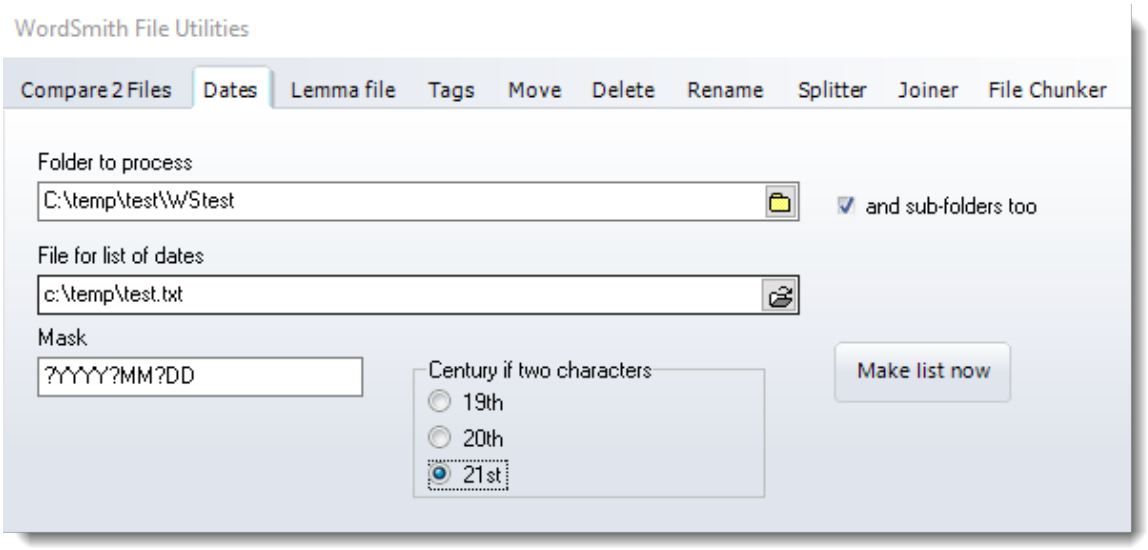
Specify your 2 files and simply press "Compare".

See also : file chunker 458, find duplicates 425, rename 451

10.5.3 dates and times

Purpose

The aim here is to parse your file-names identifying suitable textual file dates and times 149, where you have incorporated something suitable in the file-name. Suitable dates can be re-used by saving file-choices as favourites 59.



Mask Syntax

The procedure reads any file-names in the *Folder to process* (and optionally its sub-folders) and attempts to parse them. If an indicator is found it will record a suitable date combination.

Suitable indicators of textual date are

YY or YYYY	year, two or four digits (YY=a 19th, 20th or 21st Century date depending on the choice set above)
MM	month
DD	day
*	skip all characters until a digit is found

The procedure doesn't understand words such as "December" or "Five", it only uses digits. Any character other than Y,M,D,* in the mask simply gets ignored.

Show file Dates & Times

If you choose this button, the program simply examines all the text files found and lists all different file-dates and times as stored on the file system, ignoring the Mask. It lists results in the *File for list of dates*. The point of it is to check whether your texts cover the range of times & dates you expect.

Make list using Mask

The program will always add each entry found to a simple text file (*File for list of dates*) listing its file-name and adding a suitable date as expected in the auto-date procedure 56, (or **<no date found>** if the mask didn't match a valid date). In addition, where the result is 1st January 1980 or later, it will set the file's time and date in the operating system to the date as parsed, so that WordSmith will automatically match the date of the text contents to the date stored on disk.

When all files have been processed, the program opens the list of files in Notepad or equivalent. Use it afterwards in the auto-date procedure 56 within file-choosing and save your preferred text files as favourites 59.

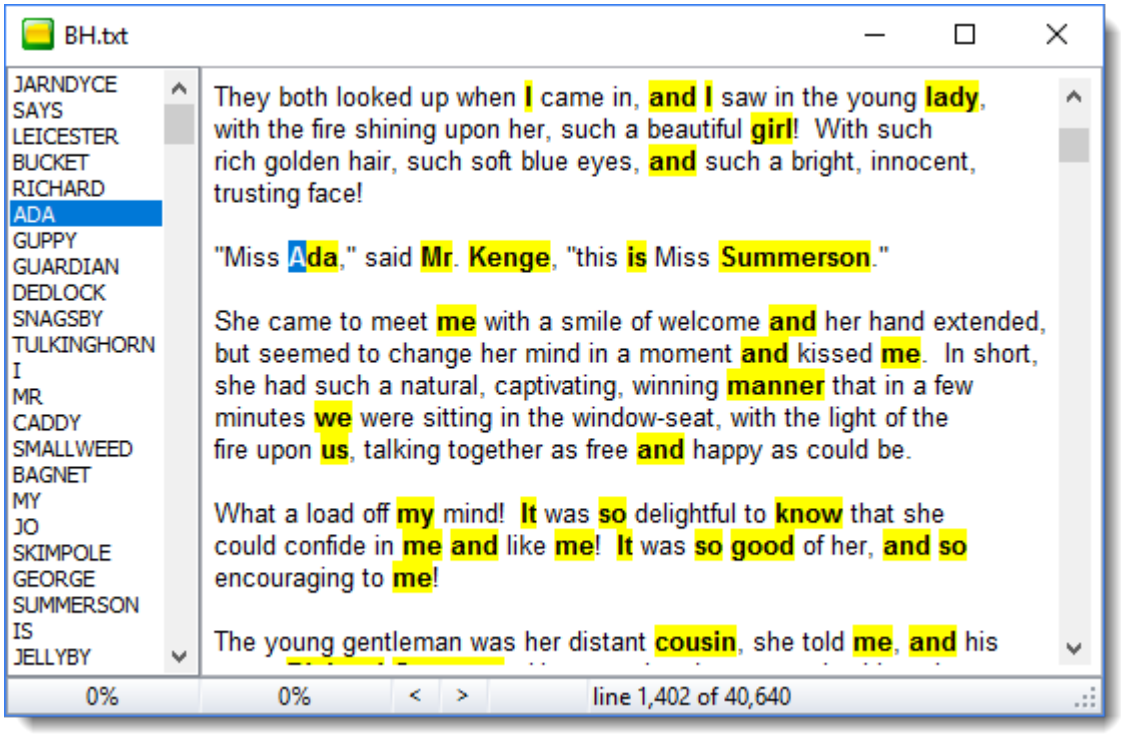
Examples

Your Mask	Source file-name	Date and Time interpreted
YYYYMMDD	20060512 Peter monologue.txt	12th May 2006 (first 8 characters used in the mask)
YYMMDD	841231.txt	31st December 1984 (20th Century assumed if YY mask used)
DDMMYY	311284.txt	31st December 1984
DDMMYYYY	20060512 Peter monologue.txt	20th June, the year 512 AD
DDMM	20060512 Peter monologue.txt	20th June of the current year
######YYYYMMDD	Peter 20060512.txt	12th May 2006 (first six characters were ignored, five for Peter, one for space)
*YYYY	Peter 20060512.txt	15 July 2006 (all characters to first digit skipped, then next 4 used for year date)
YYYY	1086 Domesday book.txt	15 July 1086 (there were only four digits)
YYYYMMDD	1086 Domesday book.txt	15 July 1086 (mask had 8 digits but file-name only 4)
YYYY#MM#DD	2006,05/12,10-54.txt	12th May 2006
YYYY MM DD	2006,05/12,10-54.txt	12th May 2006
?YYYY?MM?DD	T2015_11_27 discussion.txt	27th November 2015 (T at beginning of file-name ignored)

10.5.4 highlighter

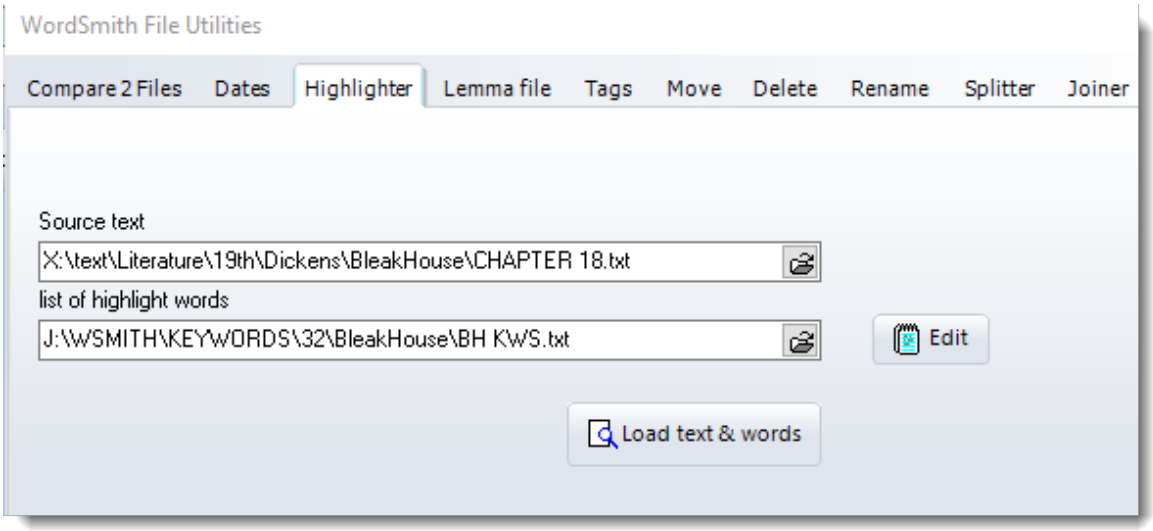
The point of it

is to show a text and selected words (or phrases) you want to highlight. Here is an example from the Dickens novel *Bleak House*, highlighting 314 key words.



How to do it...

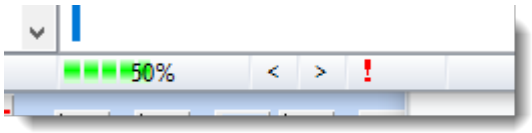
In File Utilities, choose the Highlighter tab, your text file and a plain text list of words you're interested in.



You can edit the list of words if necessary using the *Edit* button. When ready, press *Load text and words*.

Highlighting

Right-click to choose between highlighting all words or just one word you've selected at the left. Press the red exclamation in the status bar if losing patience. You can see progress marked here in the status bar.



Highlighting lots of words in a huge text takes time... finding over 300 words which each occur many times in this 800-page novel took about 6 minutes on my PC.

Trimming

Right-click to trim the lines of the text. This removes unnecessary blank lines.

Save as RTF

Right-clicking lets you save the text in Rich Text format and if you have an editor capable of showing RTF you will get to see the result.

Format of the word list

This is plain text, either one word per line or with commas separating items. Multi-word units are allowed. Searches are by default whole-word searches, case insensitive. You can alter that using the standard WordSmith search syntax 246.

The arrows

These are to take you forwards and backwards through the highlighted words. To move only from one instance of a specific word, select that word first in the left display. In the picture above, **Ada** is selected. The current selected word has a blue first letter.

10.5.5 lemma file building

The aim is to create a lemma file 327 using a corpus which has lemmas marked up.

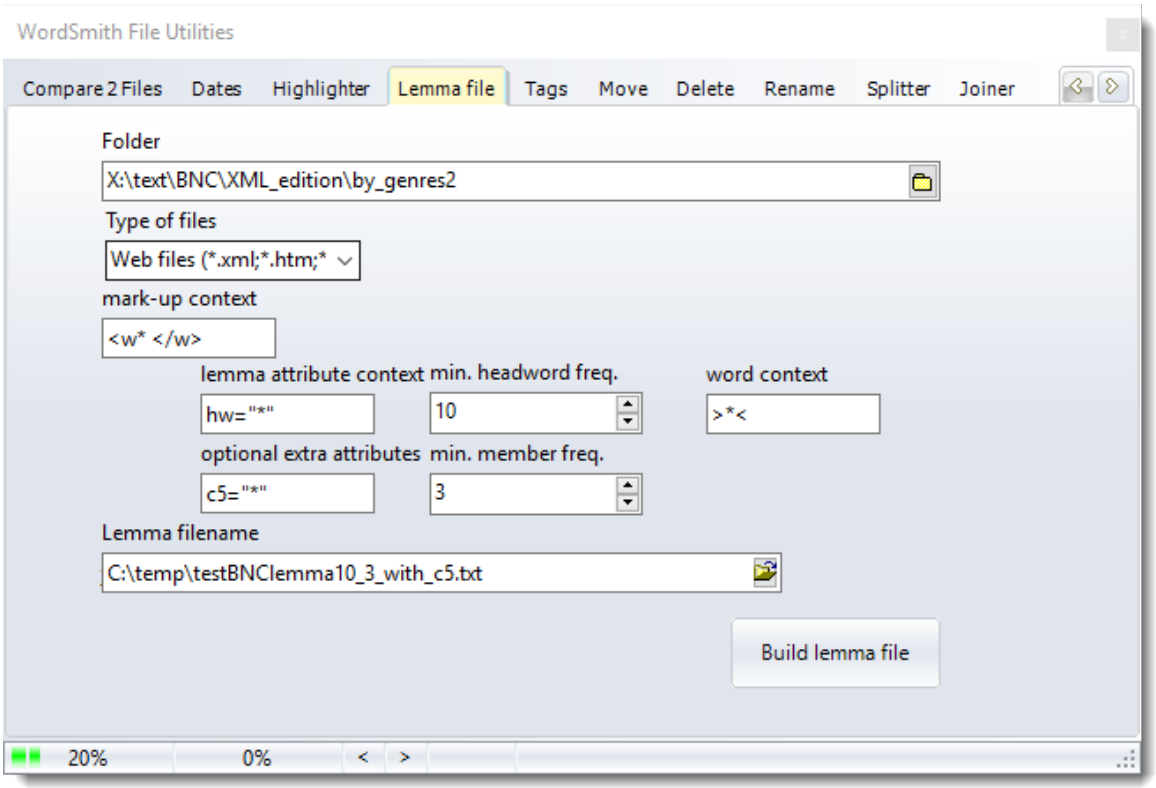
Example

The BNC XML edition uses mark-up like this:

```
.... <w c5="AJC" hw="strong" pos="ADJ">stronger </w> .....
```

where the *lemma* (head-word) is **strong** but the word in context is **stronger**. You can use the lemma file building procedure to go through all the **.xml** texts in this corpus finding all the items which are marked as belonging to the same lemma (*members* of the lemma) and saving them in a text file.

How to do it



Folder and files

Choose a folder which the corpus is headed by.This and any sub-folders will be searched for your corpus files (such as **.xml**).

Mark-up context

In the mark-up context you need to specify the start and end of each chunk of your corpus which will contain both a lemma (head-word) and an original word. The asterisk is used to represent a whole lot of characters including spaces, so for the example given above **<w* </w>** as seen in the picture is sufficient to find the mark-up, because each relevant chunk starts with **<w** and ends with **</w>**. Within that mark-up, the lemma attribute can be found by searching for **hw="*"**. (The ***** in this example would pick out the head-word **strong.**) And the word context can be defined as >*<, in other words the word **stronger** will be after a **>** and it'll run up to the following **<**.

Optional extra attributes

This lets you get results like

ACCOMMODATE -> <VVI>ACCOMMODATE,<VVN>ACCOMMODATED

In the screen-shot above, I specified the c5 attribute, which in the case of **accommodate** picked up the **VVI** and the **VVN** attributes as well as the word-forms. Leave blank if not wanted.

Minimum frequencies

This setting lets you filter out infrequent variants such as these:

ABBOTSONIAN -> ABBOTSONIANS

ACH -> ACHES

ADUMBRATE -> ADUMBRATED

The picture above shows a minimum of 10 occurrences required in the corpus for each lemma, with a minimum number of 3 members, probably more useful for a corpus of 100 million words.

Once you fill in the details and press *Build lemma file*, the process will run until it completes the task. The result will be a lemma file which may be useful in filtering word lists. Note that the accuracy of the list will depend on the BNC's own parsing in the first place, so you may find **HAVEING** as a lemma variant of **HAVE, BES** as a variant of **BE**, so the list will benefit from manual checking, especially if you set a minimum frequency of 1 for members of each lemma.

10.5.6 tags

Mis-matched tags problem

well-formed text is like this:

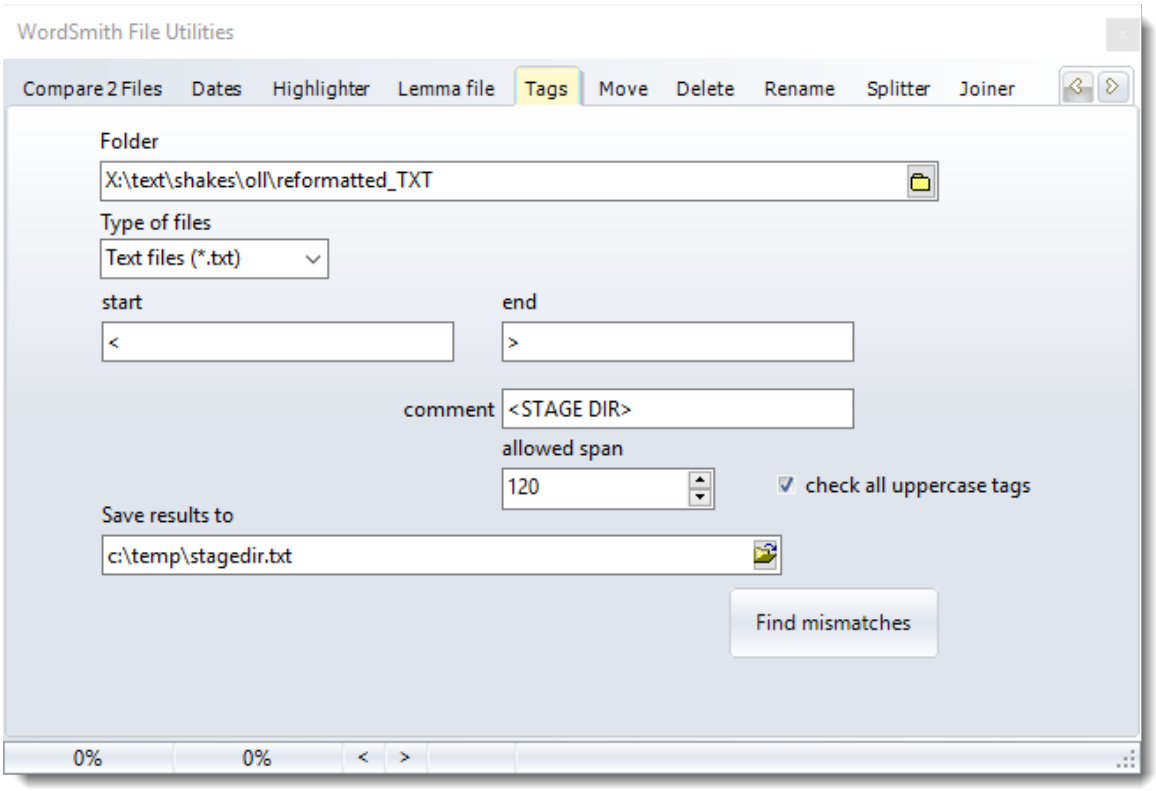
```
<introduction>blah blah blah</introduction>
<body>blah blah blah</body>
<conclusion>blah blah blah</conclusion>
```

ill-formed text might be like this:

```
<introduction>blah blah blah</introduction>
<body>blah blah blah
<conclusion>blah blah blah</conclusion>
```

The body section of the text has no explicit end. Imagine for a corpus of text files you wanted to process Only Part of Text 168, specifying that you want both the introduction and the body, avoiding the conclusion of each. If you specified that you wanted from **<introduction>** to **</body>** something would go wrong, such as the ill-formed case getting missed completely or running on to include the conclusion.

Solution



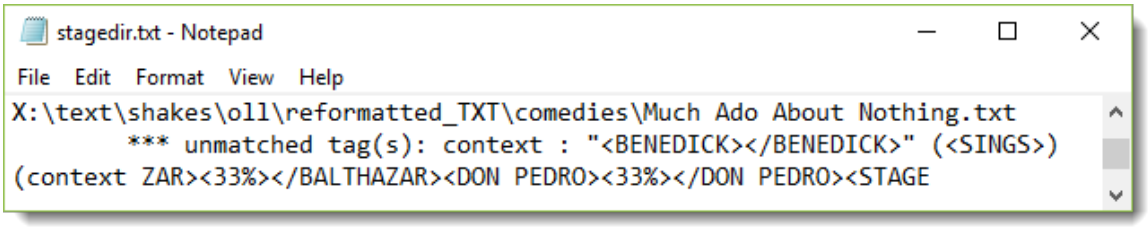
This procedure will examine all suitable text files and check whether every start has a corresponding end. Here we are checking all tags from **<** to **>**. If a tag is longer than the *allowed span* you get a warning.

The procedure checks any tag regardless of case but can optionally check all those in capitals especially thoroughly.

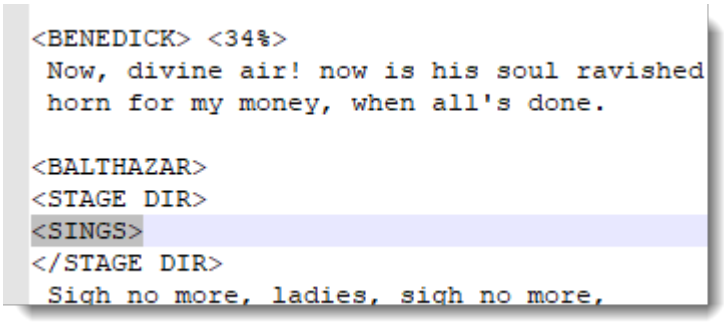
Check all uppercase tags

Checks tags if they are in capitals, looking out for mismatches (where a tag starts but doesn't end, for example) though allowing this within comment sections (here where there is a stage direction section marked **<STAGE DIR>**)

This report shows that in the play *Much Ado About Nothing*, within a section beginning **<BENEDICT>**, there is a tag **<SINGS>** which is not closed off (with **</SINGS>**). The tag context says this starts about 35% of the way through the text.



The cause of the mismatch has to be sought out. Here it is clear that **<BENEDICT>** starts but doesn't finish, because we get **<BALTHAZAR>** starting up with no sign of **</BENEDICT>**



10.5.7 move files to sub-folders

This function allows you to take a whole set of files in a folder and its sub-folders and move or copy them to suitable sub-folders elsewhere. The source file structure (patterns of folder and sub-folders) is copied for all files copied or moved.

Source files folder

Select the folder you want to find the files in.

Destination files folder

If you leave this blank the same folder will be used as the destination.

Action

copy or move files. Copying is safer!

Template use

You must specify a template identifying your needs. The first kind (explained on screen) uses a few characters to generate suitable folder names.

Example

In **c:\temp** you have

2001 Jan.txt
2001 Feb.txt
2003 Jan.txt
2003 Feb.txt
2003 March.txt
2003 Oct.txt

etc. and you want them sorted by year into different folders.

Using the template **AAAA*** you will take the first four characters of your files and place each into a sub-folder named appropriately.

Results

c:\temp\2001 contains **2001 Jan.txt, 2001 Feb.txt**

and all the others are in **c:\temp\2003**

Syntax

? = ignore this character

A = use this character in the file-name

* = use no further characters in the file-name

The second kind treats a template exactly as the file system treats file-specifications.

Example

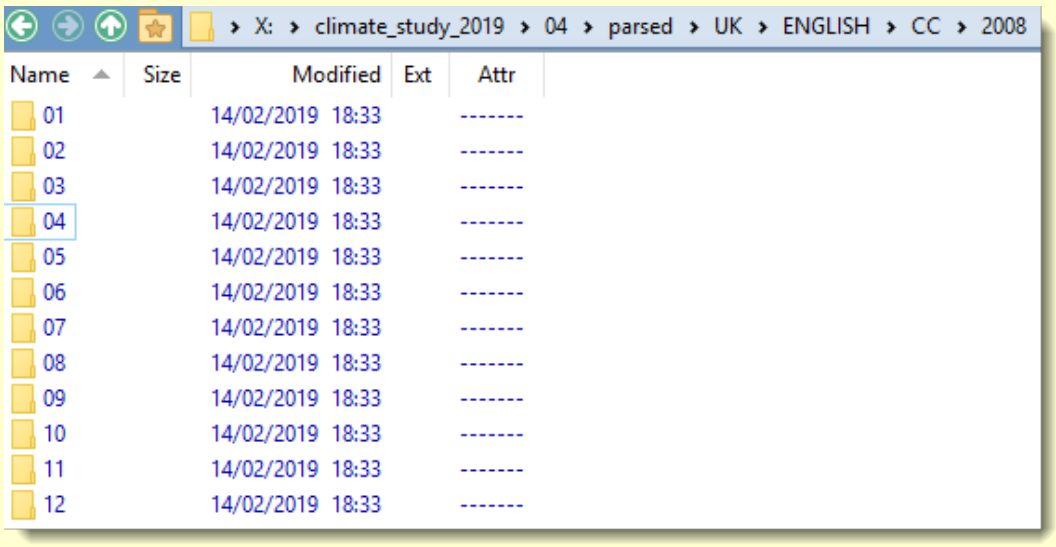
Using this template ***.txt** you will get all .txt files copied or moved.

Using **F*.txt** you will get all .txt files which start with **F** copied or moved.

With **YELLOW*.p*** you will get all files which start with **yellow** and whose extension starts with **P** copied or moved.

If you instead specify a filename, e.g. **c:\temp\my_tempates.txt** you will process all templates listed in that text file.

I wanted to copy all files beginning **0004** in a study of climate change, stored on my **J:** drive. Here you can see the structure copied to my **X:** drive year by year and month by month.

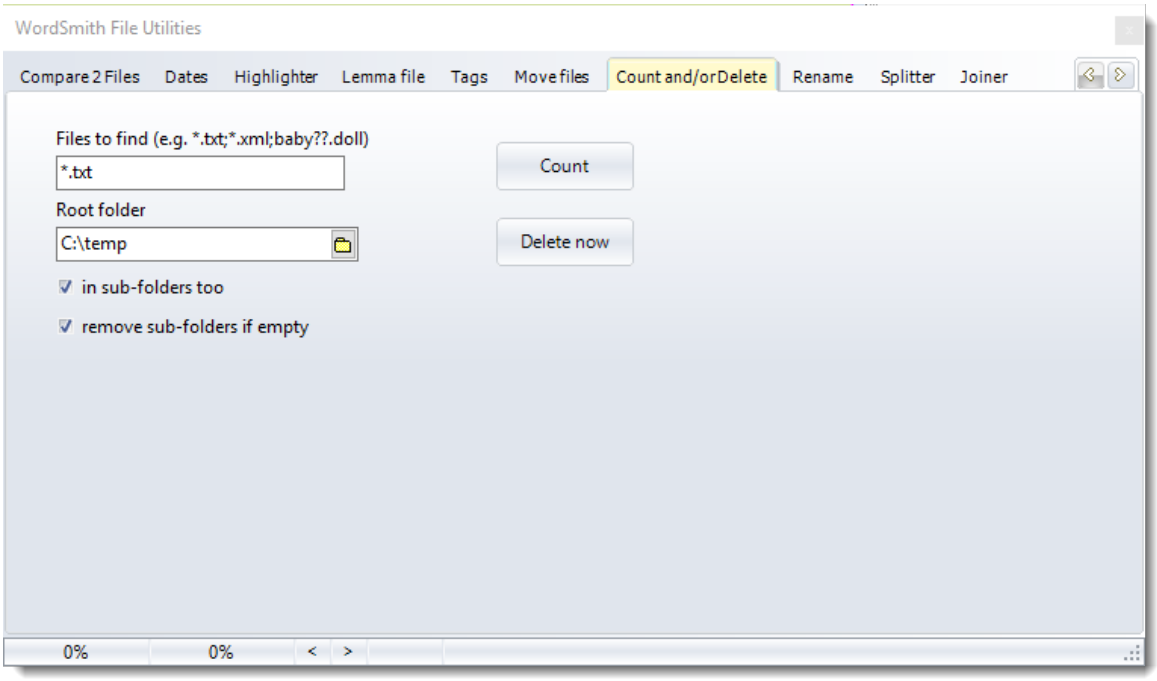


The file dates were unchanged.

10.5.8 count and/or delete

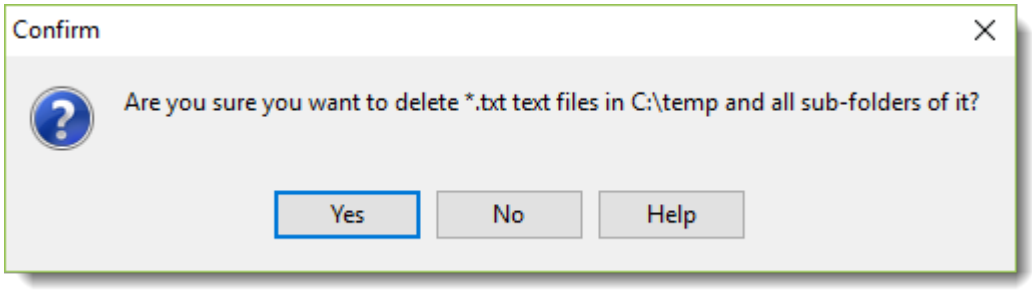
Enables you to remove text files in one sweep through your computer. Or just count how many there are...

The top button is useful if you merely want to know how many there are in a whole set of folders.



Be careful!

If you press Delete now, you will be asked for confirmation



and then the File Utilities will sweep through the root folder and optionally its sub-folders, looking for the texts specified. The example above will seek out all .txt files, then remove any sub-folders -- but only if they are completely empty.

10.5.9 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.

Template procedure:

Here you want to rename your texts so as to fit them to a template based in part on the existing file-names.

Your text files just need a date parsing and maybe some unwanted characters could be cleared too.

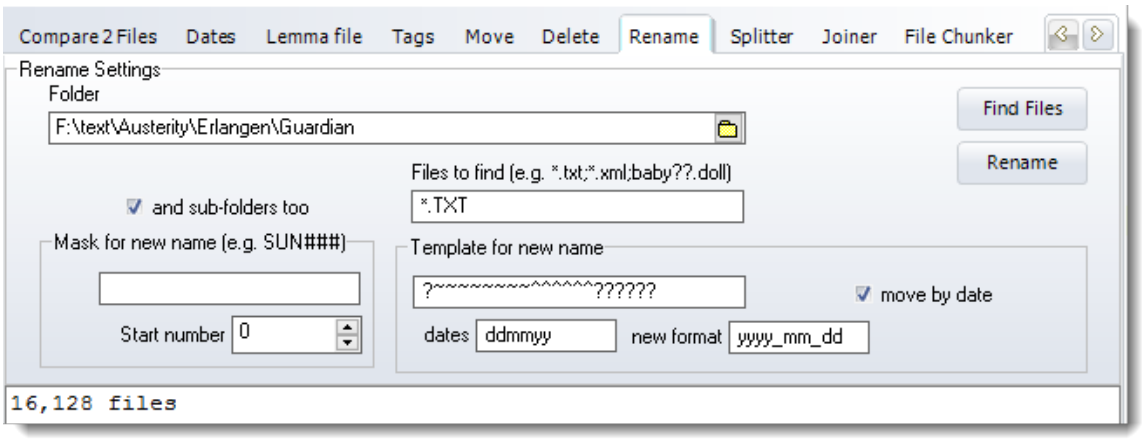
Mask procedure:

Here the idea is to rename a set of files with a standard name plus a number. For example suppose you have downloaded a lot of text files containing emails from the Enron scandal, you could rename them

Enron001.txt
Enron002.txt
etc.

How to do it

Choose the folder which your texts are in and decide whether you want to rename one in any sub-folders. Choose the types of file such as .TXT, .XML etc.



Now decide whether you want a mask procedure or a template procedure. If it's a mask and you want them all called **Enron** plus a number, just put **Enron####** where the number of # symbols shows a minimum number decimal digits in the resulting file-names. Choose a number to start the counter off with.

If it's a template, use these symbols

?	retains the existing character
~	cuts this character in the original file-name
^	these are digits which have a meaning for the date

In the example above, I wanted to keep the first character, remove the next 7, treat the following 6 as a date, and keep the next 6.

The dates in the originals were ordered **ddmmyy** (day, month, year, e.g. 310512 as 31st Oct 2012) but I wanted them re-ordered with a four-digit year (yyyy) then an underscore, then the month and day. I also wanted the texts files to be automatically dated so that their file-dates matched their publication dates.

See also : compare two files 440, file chunker 458, find duplicates 425

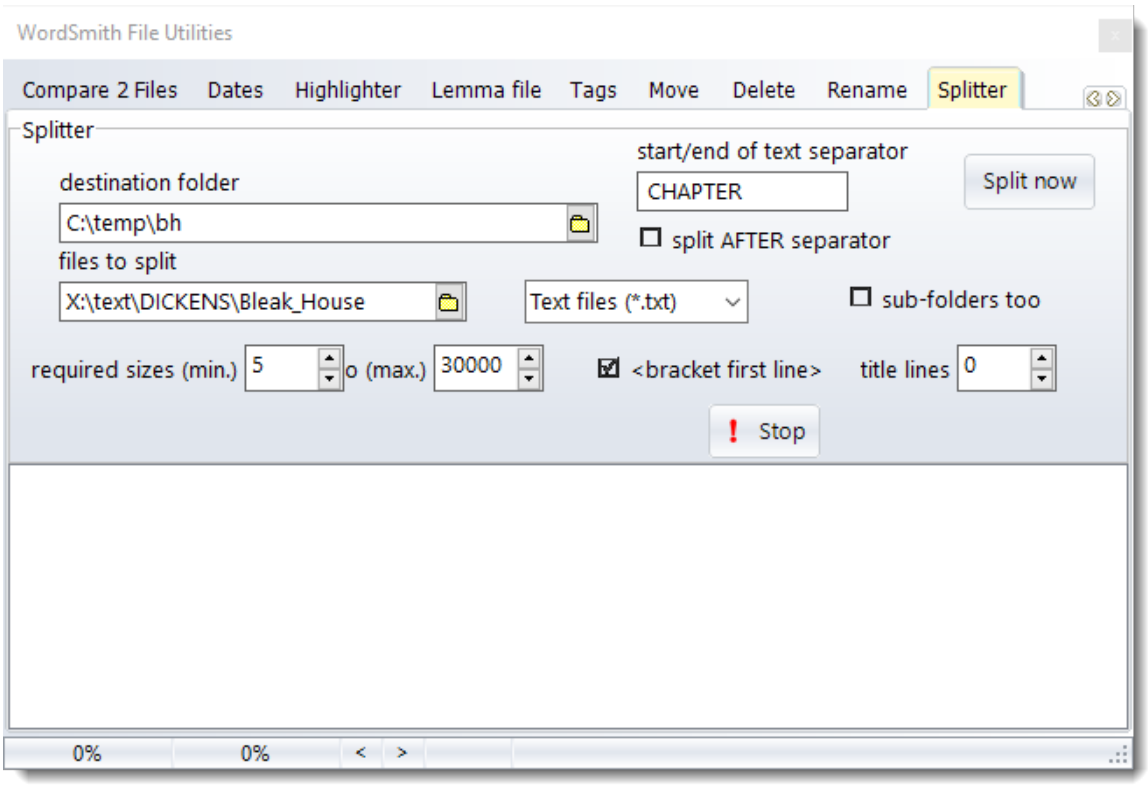
10.5.10 Splitter

10.5.10.1 Splitter: index

Explanations

This is a sub-program for splitting large files into lots of small ones. It looks in each text for a signal to show where to split the large text file and splits at that point. Here it will look for the word CHAPTER (in capitals) and if that is found will split there.

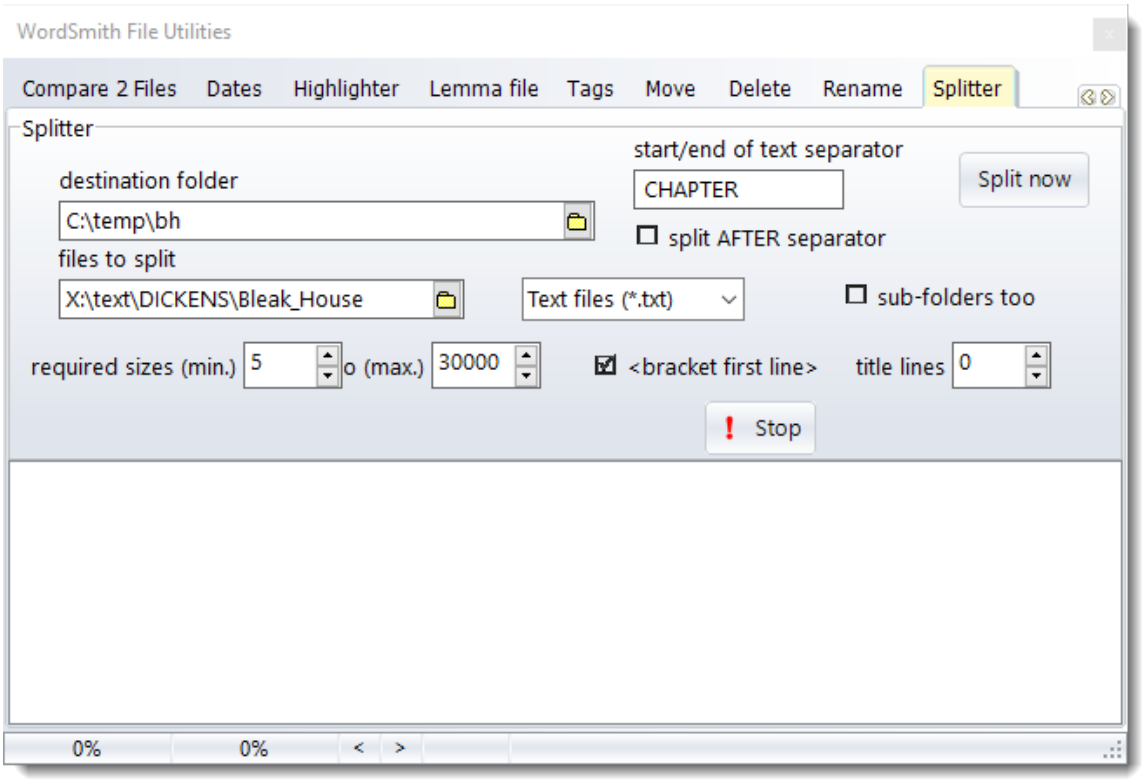
Setup 454
Filenames 456
Wildcards 457



See also : WordSmith Main Index 2

10.5.10.2 Splitter settings

Splitter needs to know how you want the large file(s) split.



start/end of text separator

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

Restrictions:

1 The start/end-of-text marker must occur at the beginning of a line in the original large file.
2 It is case sensitive: **</Text>** will not find **</text>**.
3 The first character in the separator may not be a wildcard 457 such as **#,*** or **?**.
4 ***** and # may occur only once each in the separator.

Splitter will create a new file every time it encounters the start/end-of-text marker you've specified. The *split AFTER separator* box determines whether the split occurs just when a line starting with the separator is found or just after that line.

Destination Folder

Where you want the small files to be copied to. (You'll need write permission to access it if on a network.)

Required sizes

The minimum and maximum number of lines that your small files can have (default = 5 and 30,000). Only files within these limits will be saved. This feature is useful for extracting files from very large CD-ROM files.

A "line" means from one <Enter> to the next.

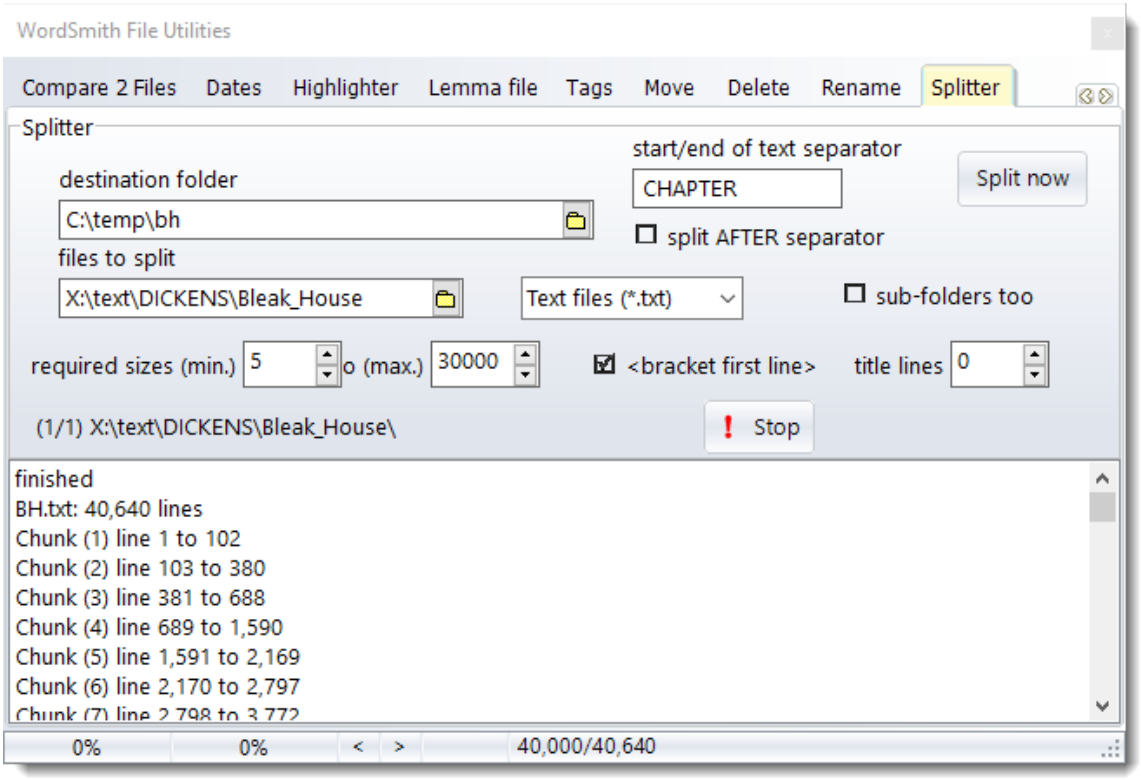
Bracket first line

Whether or not you want the first line of each new text file to be bracketed inside **< >** marks. (If your separator is a start-of-section separator like CHAPTER with a number, you may wish that to be in brackets. And often the first line after an end-of-text symbol will contain some kind of header.) If you don't want it to insert < and > around the line, leave this box unchecked.

Title Line

If you know that a given line of your texts always contains the title for the sub-text in question, set this counter to that number, otherwise leave it at 0. For example, where you know that every line immediately following **<end of story>** has a title for the next story, you could put 1.
Example :
...
<end of story>
Visiting New York
...
The file-name created for each story will contain the title as well as a suitable number. In this example a file-name might end up as **`C:\texts\split\Visiting New York 0004.txt`**.

Results

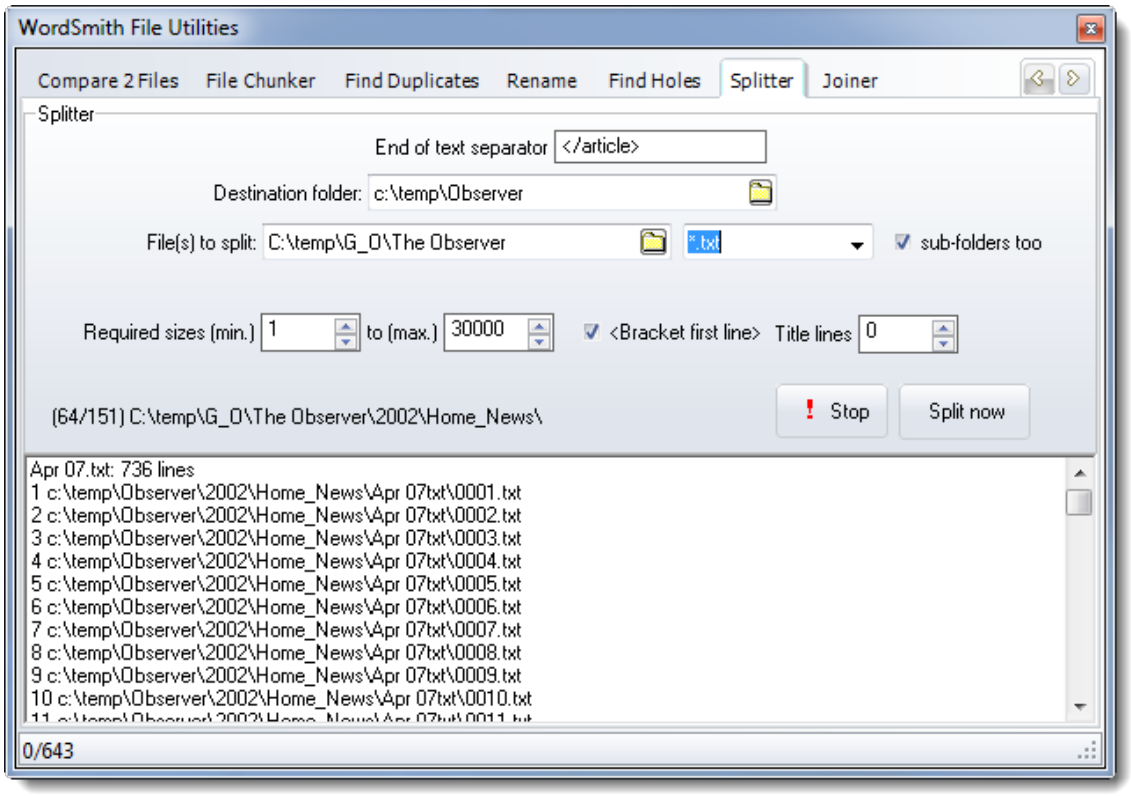


The results show that only one **.TXT** file was found (sub-folders were not searched), and it has 40,640 lines. The splits occurred at each CHAPTER.

See also: Joiner 457, Filenames 456, Wildcards 457, Text Converter index 472.

10.5.10.3 Splitter: filenames

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



It creates filenames as sub-files of a folder based on the name of each text file. In this screenshot, it has found a file called **`C:\temp\G_O\The Observer\2002\Home_news\Apr 07.txt`** and is creating a set of results listed 1 to 11 or more, using the specified destination folder plus the same folder structure as the original texts. Each sub-text is numbered **`0001.txt, 0002.txt`** etc.

Sub-folders are created if there are too many files for a folder.

If a title is detected, each file will contain the title plus a number and .txt. If there is no title, the filename will be the number + .txt added as a file extension.

Tips

1. Splitter will start numbering at 1 each session.
2. Note that the small files will probably take up a lot more room than the original large file did. This is because the disk operating system has a fixed minimum file size. A one-character text file will require this minimum size, which will probably be several thousand bytes in size. Even so, I suggest you keep your text files such that each file is a separate text, by using Splitter. When doing word lists and key words lists, though, do them in batches 30.
3. CD-ROM files when copied to your hard disk may be read-only. You can change this attribute using Text Converter 472.

10.5.10.4 Splitter: wildcards

- **#** The hash symbol, #, is used as a wildcard to represent any ***number***, so **[FF#]** would find *[FF3]* or *[FF9987]* but not *[FF]* or *[FF 9]* (because there's a space in it) or *[FFhello]*.
- ***** The asterisk represents any ***string***, so **[FF*** would find all of the above. * is used as the last character in the end-of-text symbol. It would find *[FF anything at all up to the next <Enter>*.
- **^** The ^ mark represents any single ***letter***, so **[FF^^]** would find *[FFZQ]* but none of the others.
- **?** The question mark represents any single ***character*** (including spaces, punctuation, letters), so **[FF??]** would find *[FF 9]* in the above examples, but none of the others.

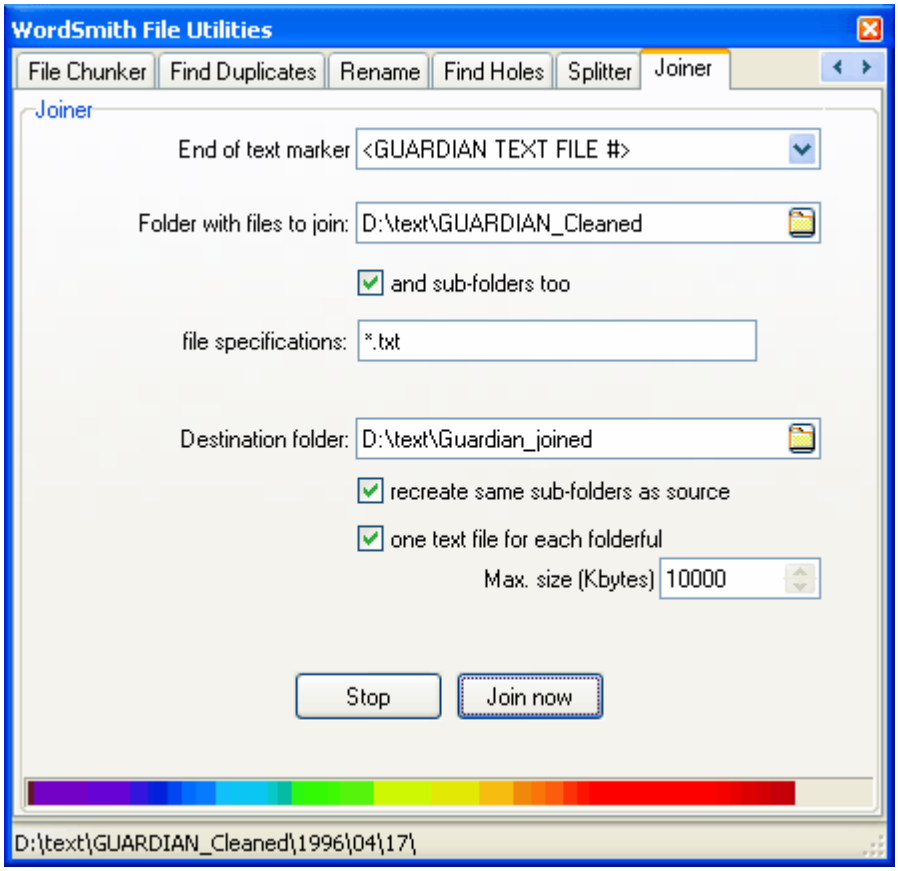
To represent a genuine **#,^,?** or *****, put each one in double quotes, eg. **"?" "#" "^" "*"**.

See also: Settings 454, Wildcards 148

10.5.11 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 in the source. In the example, all the sub-folders of **`d:\text\guardian_cleaned`** will be created below **`d:\text\guardian_joined`**.

one text for each folderful

if checked, a whole folderful of source texts will go into one text file in the destination.

Max. size (Kbytes)

The maximum size in kilobytes that you want the each merged text file to be. 1000 means you will get almost 1 megabyte of text into each. That is about 150,000 words if there are no tags and the text is in English. This only applies if *one text for each folderful* isn't checked.

Stop button

Does what it says on the caption.

See also: Splitter 453, Text Converter index 472.

10.5.12 file chunker

The point of it

The idea is to be able to cut up a big file into pieces, so that you can copy it in chunks e.g. for emailing.

Naturally, you may later want to restore the chunks to one file.

How to do it: to copy a file

1. Specify your "file to chunk" (the big one you want to copy)
2. Specify your "drive & folder" (where you want to copy the chunks to.
3. Specify the "size of each chunk"

4. Specify whether to "compress while chunking" (compresses the file as it goes along)
5. Press "Copy".

How to do it: to restore a file

1. Specify your "first chunk" (the first chunk you made using this program)
2. Specify which folder to "restore to" (where you want the results)
3. Specify whether to "delete chunks afterwards" (if they are not needed)
4. Press "Restore".

See also : compare two files 440, find duplicates 425, rename 451

10.5.13 find holes in texts

After text files have been copied from one source to another, they may get slightly corrupted with holes 574 in the stream of text. This utility lets you seek out the texts in your corpus which have got corrupted in this way and optionally lets you delete them. If you want to convert the holes to space-characters, use the Text Converter.

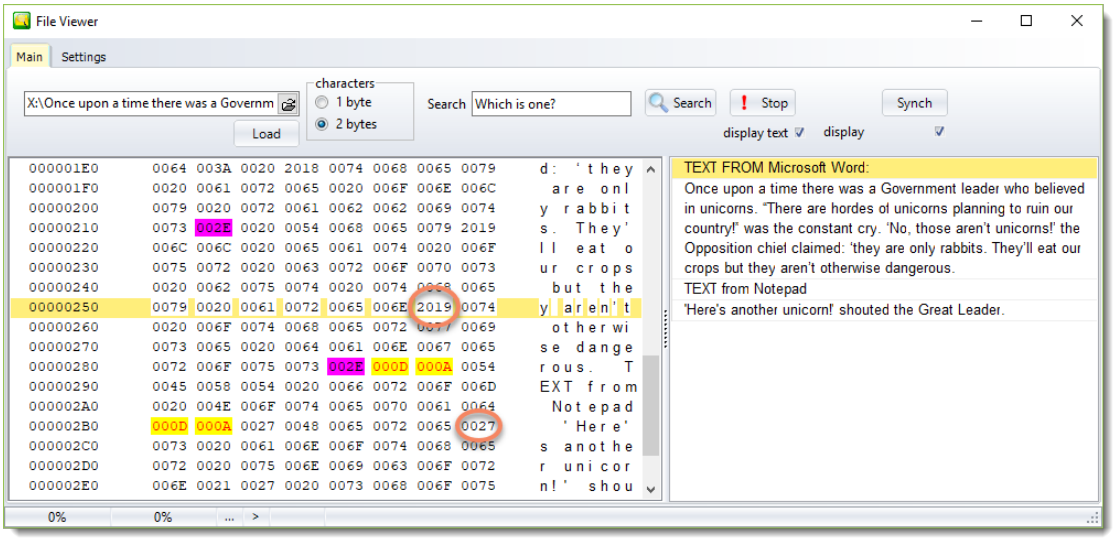
10.6 File Viewer

10.6.1 Using File Viewer

Aim

To help you examine files of various kinds to see what is in them. This might be

- in order to see whether they're really in plain text format
- to see whether there's something wrong with them, such as unusual characters which oughtn't to be there
- to see whether they were formatted for Windows, Mac, or for Unix
- to check out any hidden stuff in there. (A **Word .doc** 552 for example will have lots of hidden stuff you don't see on the screen but is inside the file anyway, such as the name of the author, type of printer being used, etc.)
- to find strings of words in a database, a spreadsheet or even a program file.
- to get certain selected characters picked out in an easy-to-find colour



Here you can see the gory details of the text. Some characters are highlighted in different colours so you can see exactly how the text is formatted. The purple 002E is the full stop. Character 2019 is the right-curling apostrophe in a section written using Microsoft Word. 0027 is an ordinary apostrophe used three times in the Notepad section.

Loading a text file

Choose your file – if necessary click on the button at the right of the text-input box. Press *Show*. Or simply drag a file from another program onto this one.

Characters

The two options available are as 1 bytes or 2 to represent each character-symbol in the text in question. You may need to alter this setting to see your text in a readable format.

The two windows

The left window shows how the "text" is built up. You can see each character as a number and, further to the right, as a character.

The right window shows the text, paragraph by paragraph, word-wrapped so you can read it.

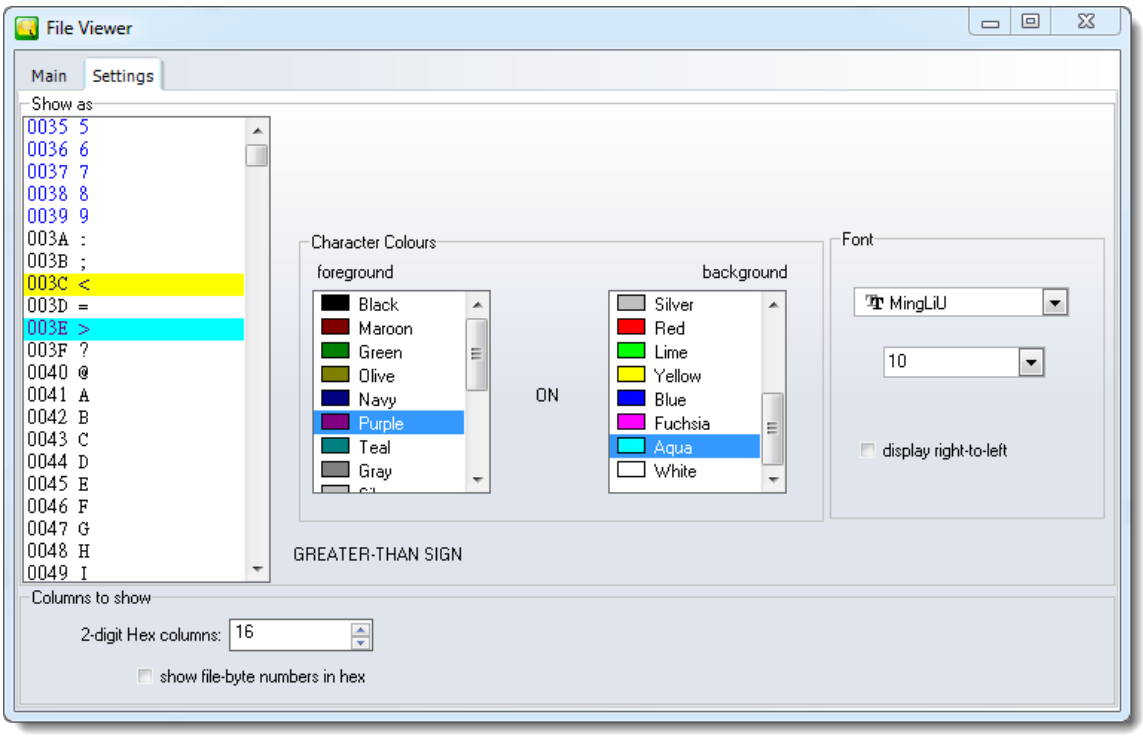
Searching

Just type in the search-word and press Search. The search is case sensitive and is not a "whole word" search.

Synch

Press the *Synch* button to synchronise the two displays. The display you clicked last is the "boss" for synchronising.

Settings



Colours

The colour grids let you see the number section in special colours, so you can find the potential problems you're interested in.

- First select the character you want coloured.
- Click the foreground or background colour list change the colour.

The character names are Unicode names. In the picture the symbol with the 003E code (>) is the last one clicked.

Font

Choose the font and size in the font window. You may need to change font if you want to see Chinese etc. represented correctly.

Columns

- You can set the "hex" columns between 2 and 16.
- You can see the numbers at the left of the main window in hex or decimal.

See also: Character Profiler 404, characters within a word 146

10.7 Minimal Pairs

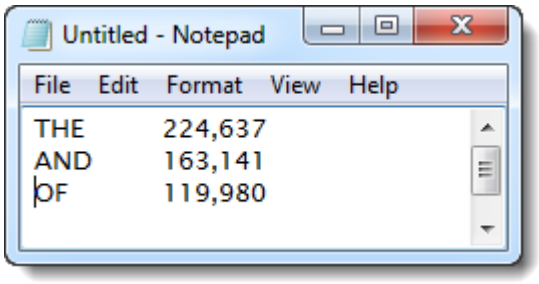
10.7.1 aim

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.

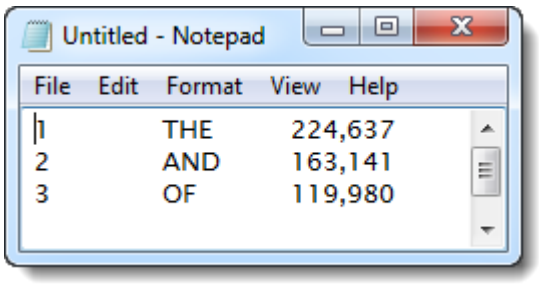
See also : requirements 462, choosing your files 464, output 466, rules and settings 467, running the program 468.

10.7.2 requirements

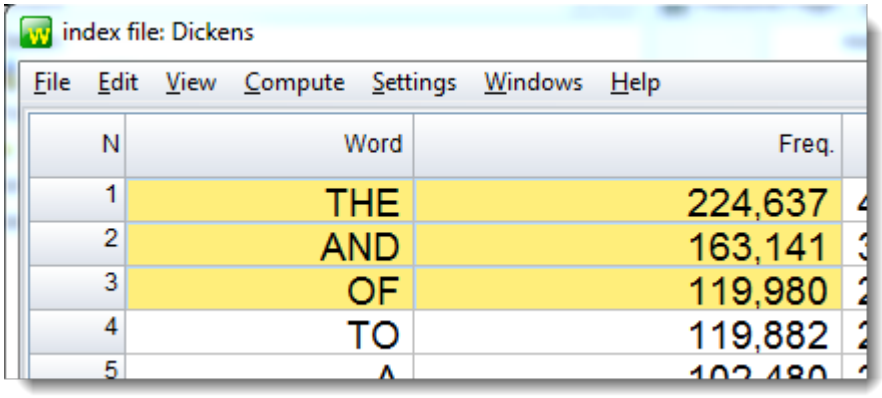
A word-list in text format. Each line should contain a word and its frequency separated by tabs, e.g.



or

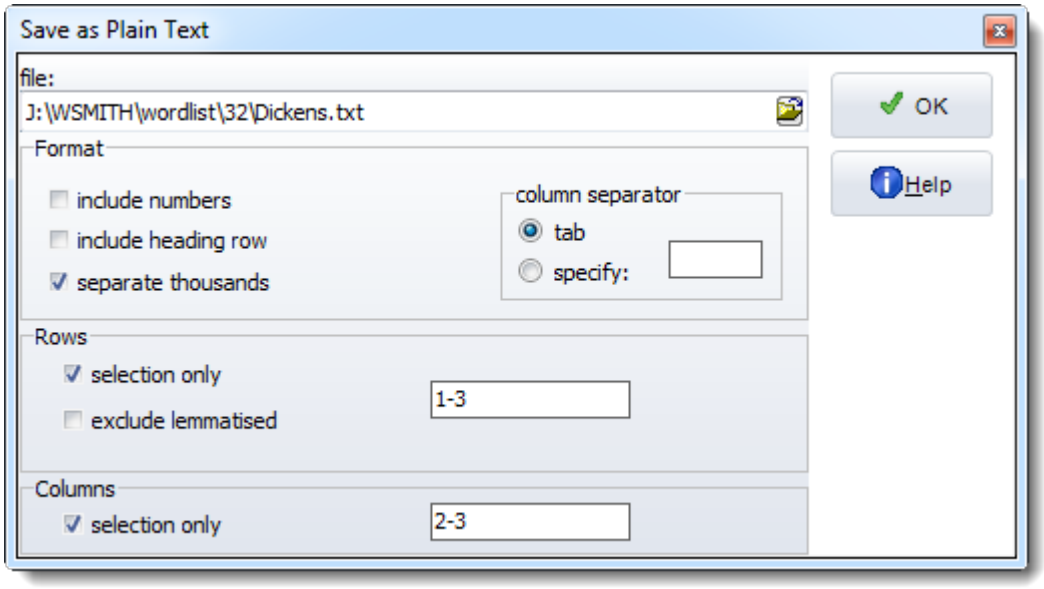


You can make such a list using WordList 7. For example, select (highlight) the columns containing the word and its frequency and

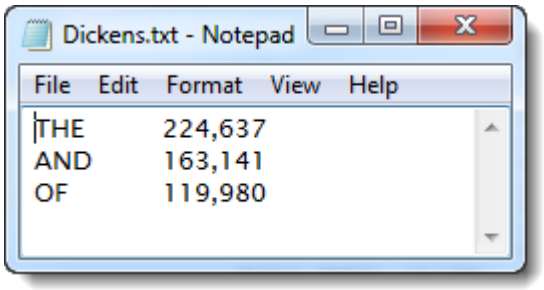


copy to the clipboard, then paste into Notepad, or

save as TXT (without numbers or heading row):



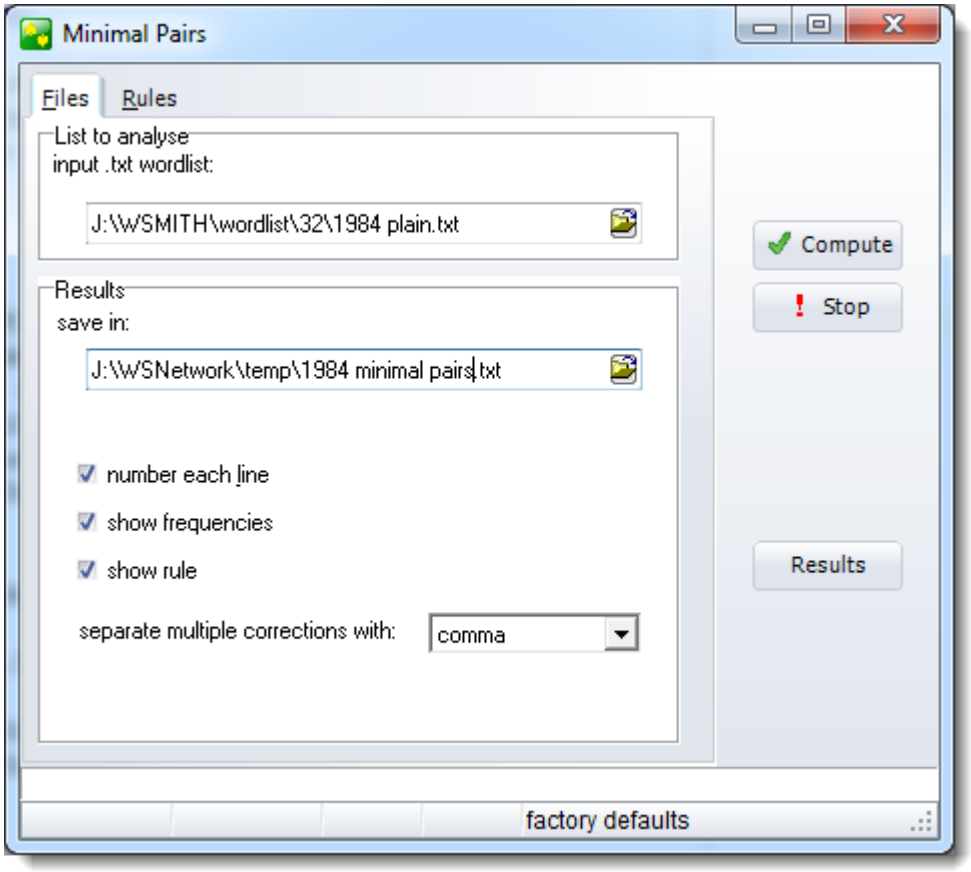
giving this:



See also : aim 462, choosing your files 464, output 466, rules and settings 467, running the program 468 .

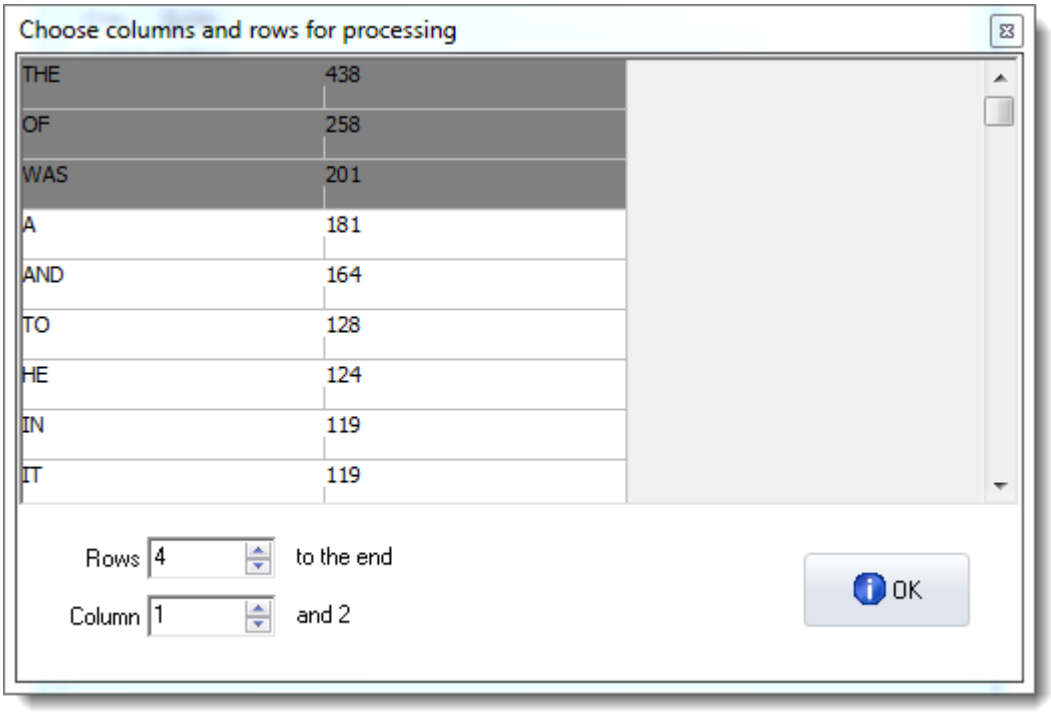
10.7.3 choosing your files

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.

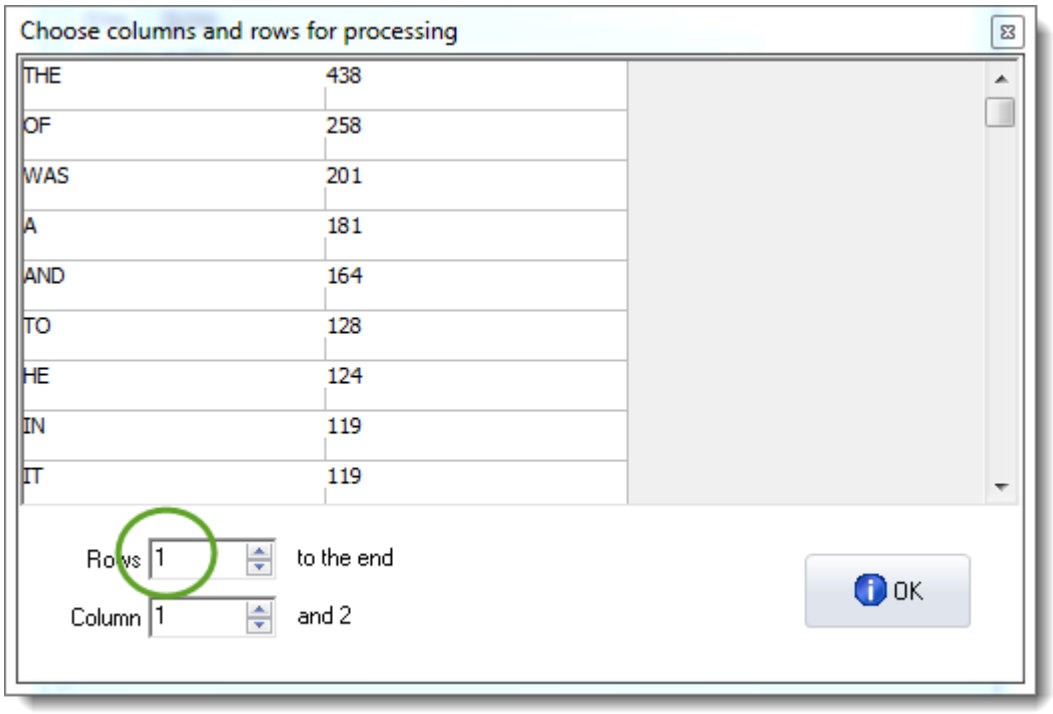


Type in an appropriate file-name for your results. Choose the rules 467 too. When you're ready, press the *Compute* button.

You'll then be asked to choose the columns and rows (allowing you to skip header lines or the number column if your txt file has those).



Here, the first three lines are greyed out, so we need to alter the Rows box:



See also : aim 462, requirements 462, output 466, rules and settings 467, running the program 468.

10.7.4 output

An example of output is

```
418    ALTHOUGHT      (7)     ALTHOUGH(37975)
```

Here the lines are numbered, and the bracketed numbers mean that ALTHOUGHT occurred 7 times and ALTHOUGH 37,975 times.

An example using Dutch medical text, lower case:

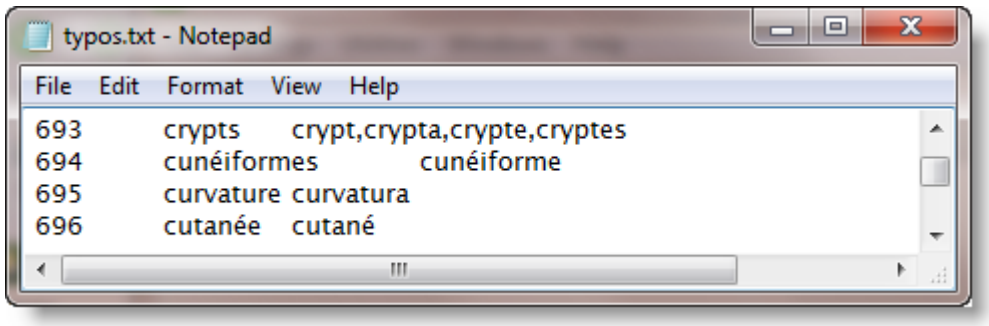
```
136    aplasie        (1)     aplasia(1)[L]
137    apyogene       (1)     apyogeen(1)[S]
138    arachnoideales         (1)     arachnoidales(1)[I]
```

Here line 136 generated a 1-Letter difference, 137 a Swap and 138 an Insertion.

An example using Guardian newspaper, looking for anagrams:

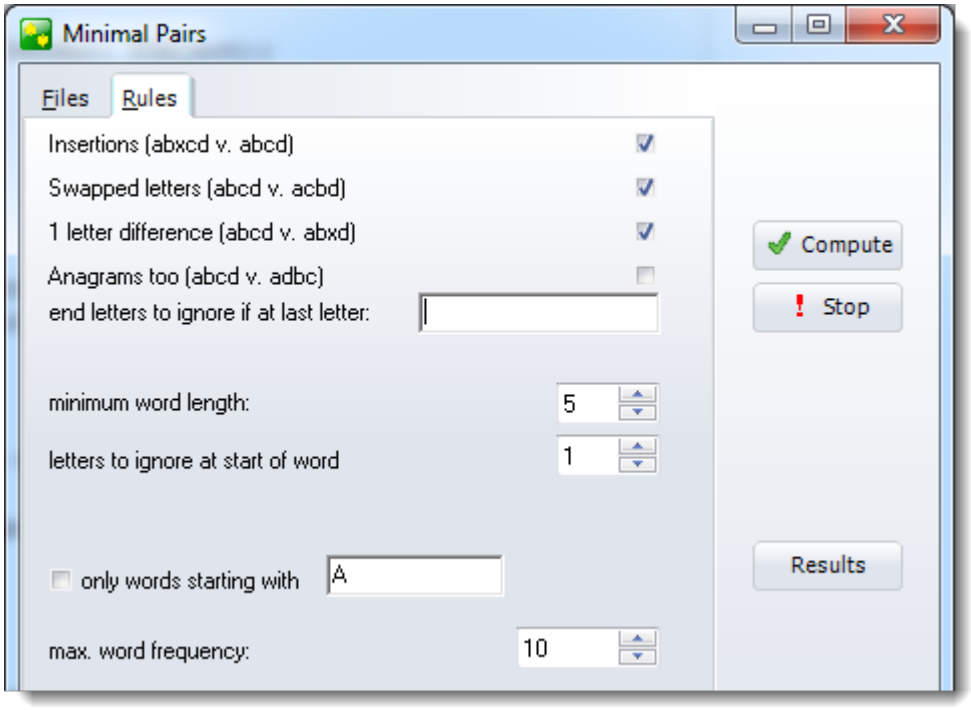
```
35     AUDIE  (7)     ADIEU(43)[A]
36     ABASS  (6)     ASSAB(16)[A]
37     AGUIAR (6)     AURIGA(11)[A]
38     ALRED'S        (6)     ADLER'S(18)[A]
39     ANDOR  (6)     ADORN(128)[A]
```

an example where the alternatives are separated with commas but the rule and frequencies are not shown.



See also : aim 462, requirements 462, choosing your files 464, rules and settings 467, running the program 468.

10.7.5 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 word length

This setting specifies the minimum word length for the program to consider the possibility there is a typo. The default is 5, which means 4-letter words will be simply ignored. This is to speed up processing, and because most typos probably occur in longer words.

letters to ignore at start of word

This setting (default =1) allows you to assume that when looking for minimal pairs there is a part of each at the beginning which matches perfectly. For example, when considering ALEADY, the program probably doesn't need to look beyond words beginning with A for minimal pairs. If the setting is 1, it will not find BLEADY as a minimal pair. To check all words, take the setting down to 0. The program will be 26 times slower as a result!

only words starting with …

If you choose this option, the program will ignore the next setting (max. word frequency). Here you can type in a sequence such as F,G,H and if so, the program will take all words beginning F or G or H (whatever their frequency) and look for minimal pairs based on the rules and settings above.

max. word frequency

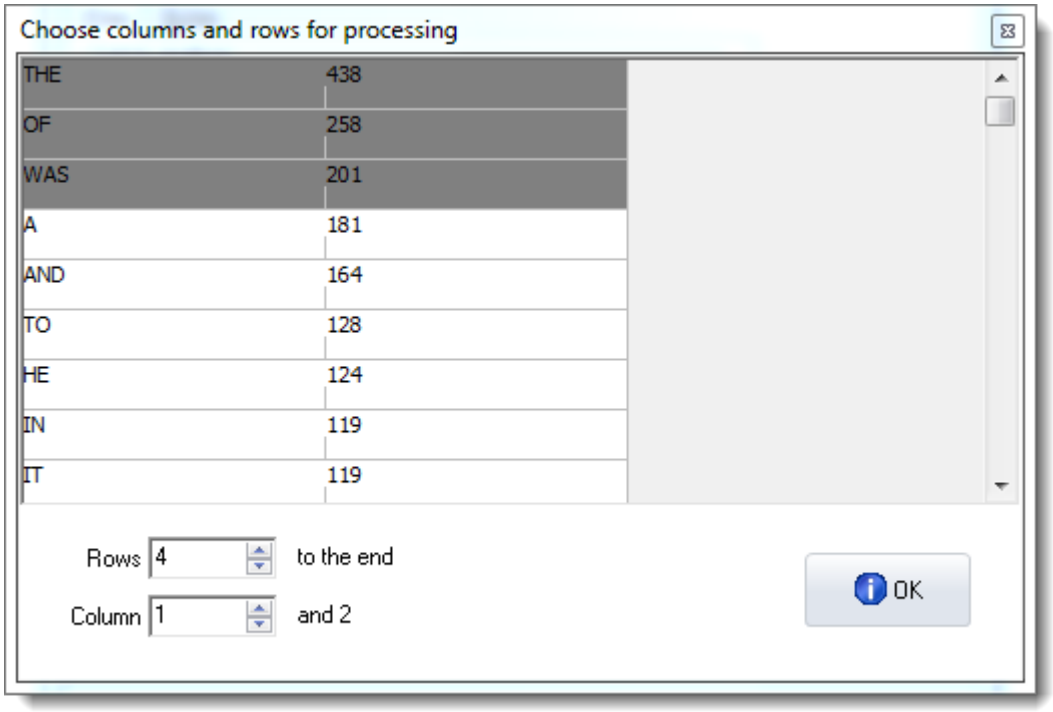
(ignored if "all words starting with" is checked) How frequent can a typo be? This will depend on how much text your word-list is based on. The default is 10, which means that any word which appears 11 times is assumed to be OK, not a typo.

Factory Defaults (restores default values)

See also : aim 462, requirements 462, choosing your files 464, output 466, running the program 468.

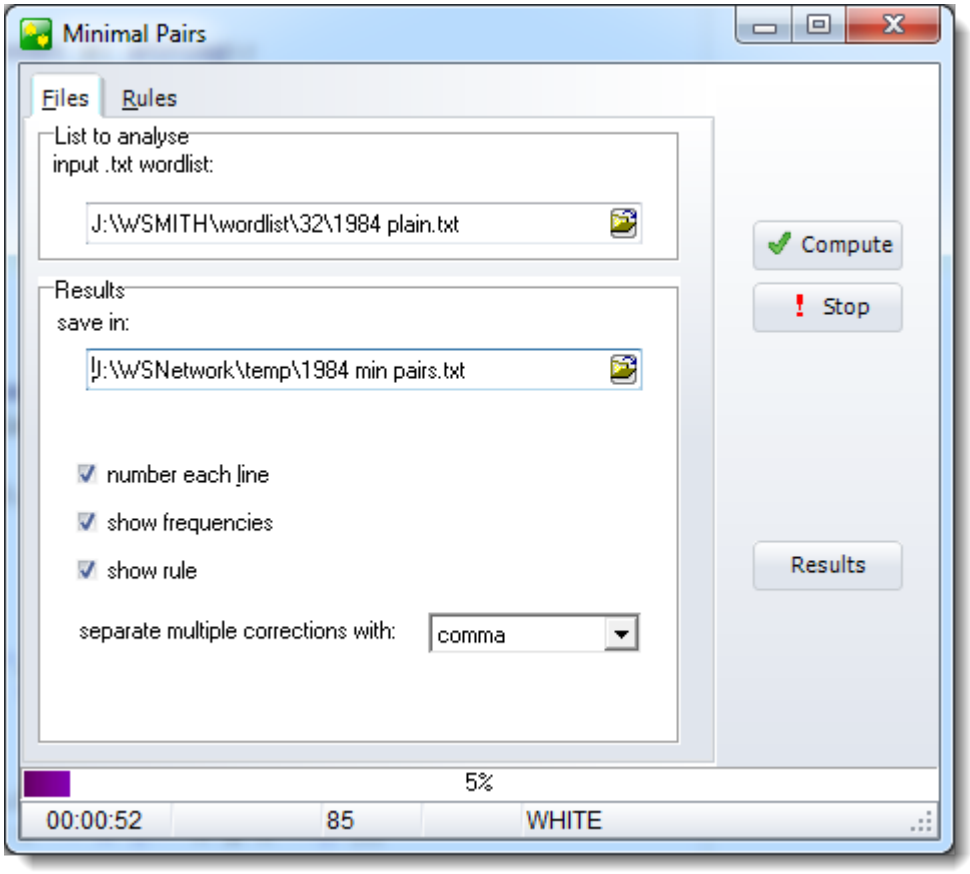
10.7.6 running the program

- 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.



Here the first three lines are greyed out, and that can be fixed by changing Rows from 4 to 1.

Once you press OK the program starts:



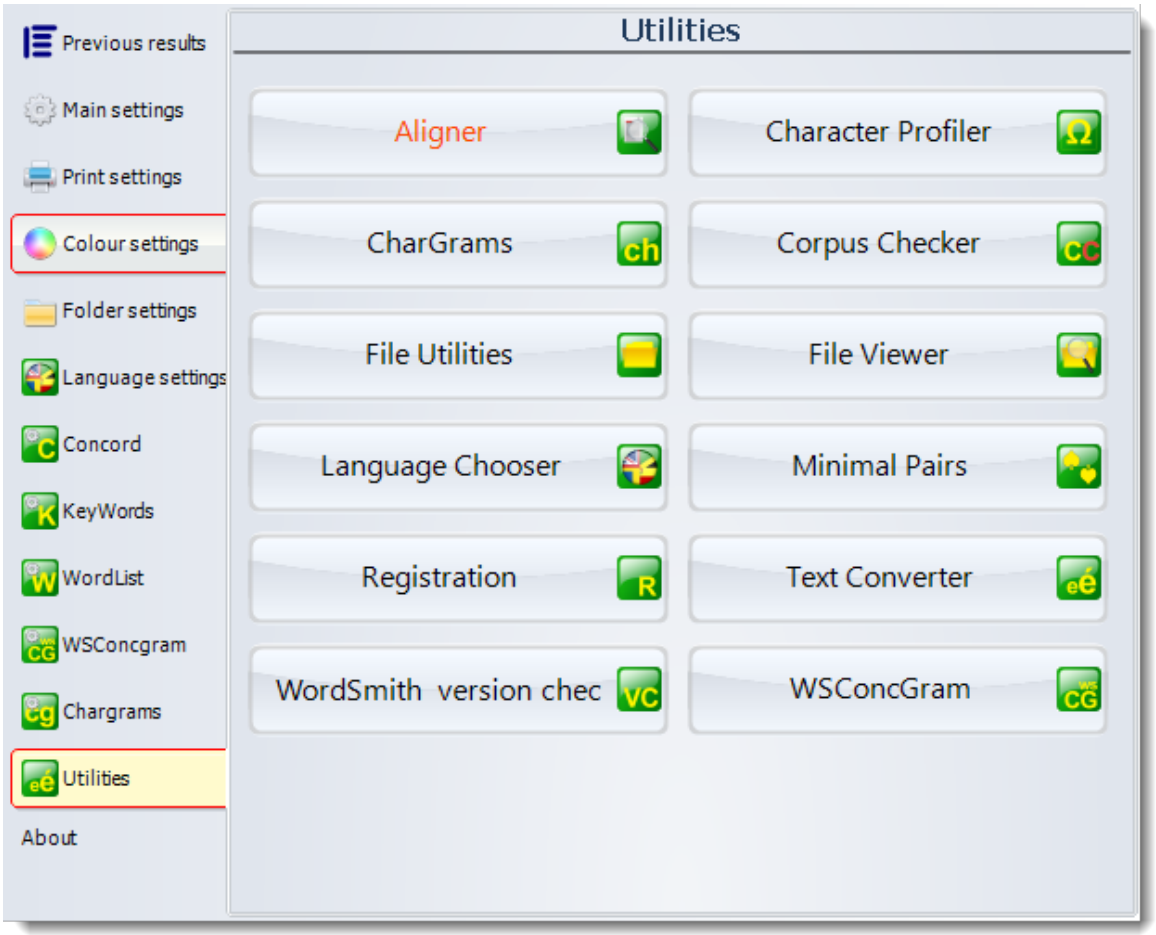
If you want to stop in the middle, press "Stop".

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

See also : aim 462, requirements 462, choosing your files 464, output 466, rules and settings 467

10.8 Registration

The Registration tool is found in the Utilities tab:



It opens automatically for unregistered users.

See also: Installation 22, Licence 24

10.9 Text Converter

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 484 so that **Text Converter** knows which symbols or strings to convert. It operates under Windows and saves using the Windows character set 527, but will convert text using DOS or Windows character sets. You can use it to make your text files suitable for use with your Internet browser.

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

Once the conversion file is prepared and Settings 473 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 473.

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 479 .

Tip

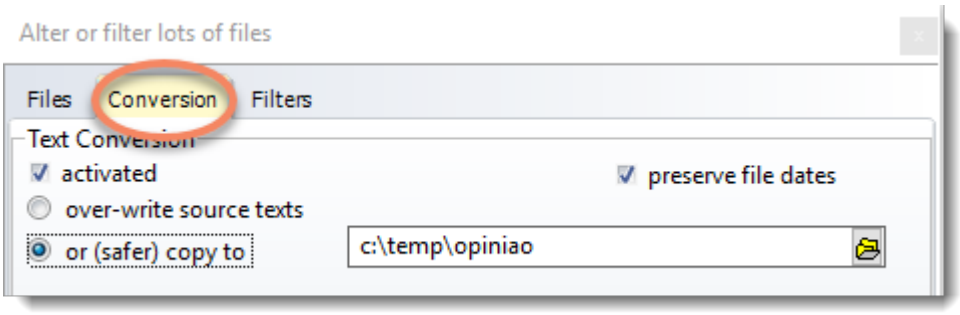
The easiest way to ensure your text files are the way you want, especially if you have a very large number to convert, is to copy a few into a temporary folder and try out your conversion file with the Text Converter. You may find you've failed to specify some necessary conversions. Once you're sure everything is the way you want it, delete the temporary files.

See also: Text Converter Contents 472.

10.9.1 index

Choose your text files

Convert text files



Filter text files into different folders

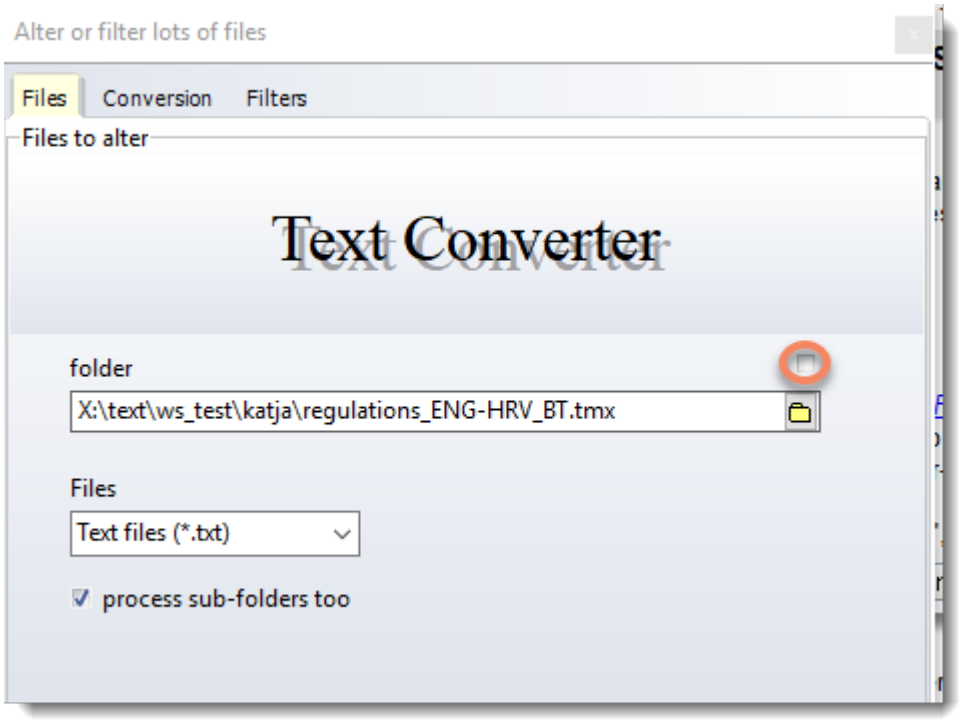
Explanations

What is the Text Converter and what's it for? 471
Getting Started... 473
Convert the text format 489
Filters 479
Sample Conversion File 486
Syntax 485
Conversion File 484

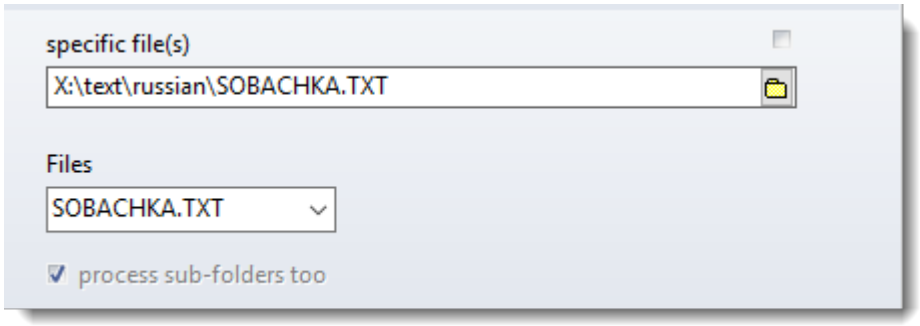
See also : WordSmith Main Index 2

10.9.2 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.

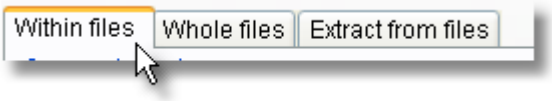


If you want one specific file, click the small checkbox with the red circle till it offers *specific file*.



2. Click on the *Conversion* or *Filters* 479 tab, and:
3. Decide whether you want to make copies of the text files, or to over-write the originals. Obviously you must be confident of the changes to choose to over-write; copying however may mean a problem of storage space.

Choose between "Within files", "Whole files" or "Extract from files"



Within files = make some alterations to specific words in each text file, if found
For example, specify what to convert, that is the search-words and what you want them to be replaced with. For a quick conversion you can simply type in a word you want to change and its replacement (e.g. *Just one change* so that **`responsable`** becomes **`responsible`**) or you can choose your own pre-prepared Conversion File 484.

Whole files = make some alterations affecting all the words in each text file
E.g. in the Whole Files section you can choose simply to update legacy files 489 in various ways, e.g. by choosing

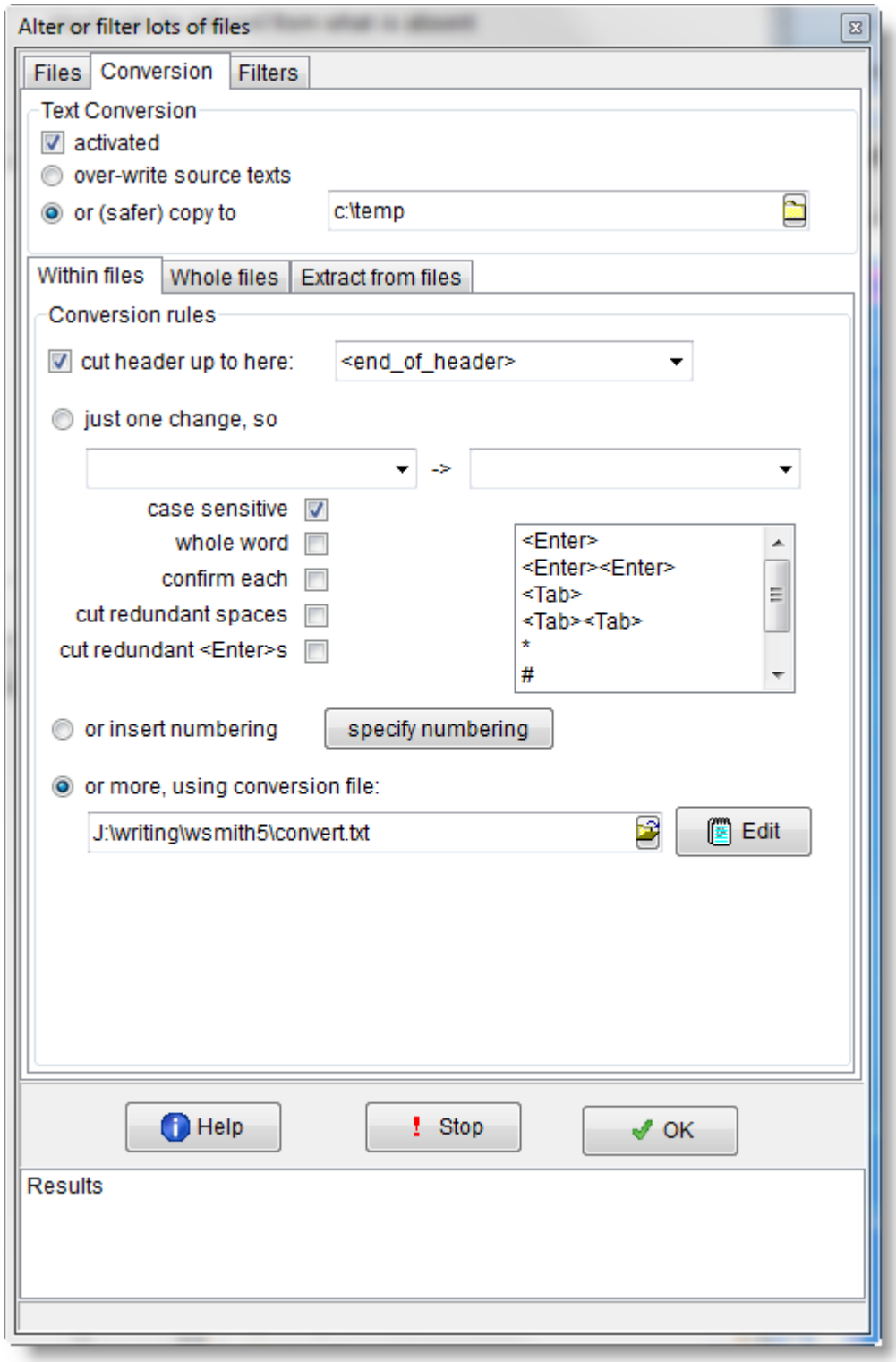
Dos to Windows,
Unix to Windows,
MS Word .doc to .txt,
into Unicode,
etc.

Or if you want simply to extract some text from your files, you should choose the Extract from files 478 tab.

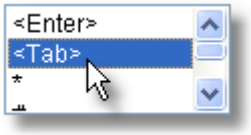
If you might want some files not to be converted, or simply don't want any conversions but instead to place files in appropriate sub-folders, choose the Filters 479 tab at the top.

If you choose *Over-write Source texts,* Text Converter will work more quickly and use less disk space, but of course you should be quite sure your conversion file codes are right before starting! See copy to 477 for details of how the folders get replicated in a copy operation.

Note that ***some space on your hard disk will be used even if you plan to over-write***. The conversion process does its work, then if all is well the original file is deleted, and the new version copied. There has to be enough room in the destination folder for the largest of your new files; it is much quicker for it to be on the same drive as the source texts. If it isn't, your permission will be asked to use the same drive.



inserting <Tab>, <Enter> etc

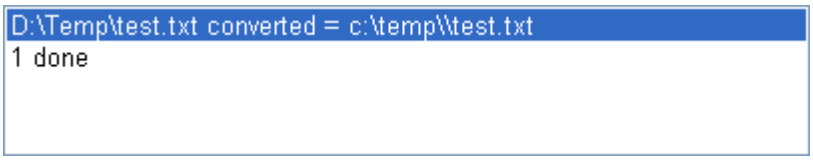


Choose in the listbox and drag to one of the windows to left or right of ->. The string inserted will conform to the format 484.

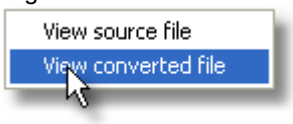
cutting out a header from each file

It can be useful to get a header removed. In the screenshot example, any text which contains **</teiHeader>** will get all the beginning of the file up to that point cut out.

Press *OK* to start; you will see a list of results, as in the screenshot below.
If you want to stop **Text Converter** at any time, click on the Stop button or press Escape.



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



See also: Text Converter Contents 472.

10.9.2.1 Text Converter: copy to

If you choose to copy the files you are converting, instead of converting or filtering them in place, which is a lot safer, the new files created will be structured like this.

Suppose you are processing **d:\texts\2007\literature** and copying to **c:\temp**
and suppose **d:\texts\2007\literature** contains this sort of thing:

```
d:\texts\2007\literature\shakespeare\hamlet.pdf
d:\texts\2007\literature\shakespeare\macbeth.pdf
...
d:\texts\2007\literature\shakespeare\poetry\sonnet1.pdf
d:\texts\2007\literature\shakespeare\poetry\sonnet2.pdf
...
d:\texts\2007\literature\french\victor hugo\miserables.pdf
d:\texts\2007\literature\french\poetry\baudelaire\le chat.pdf
...
```

you will get

```
c:\temp\shakespeare\hamlet.txt
c:\temp\shakespeare\macbeth.txt
...
```

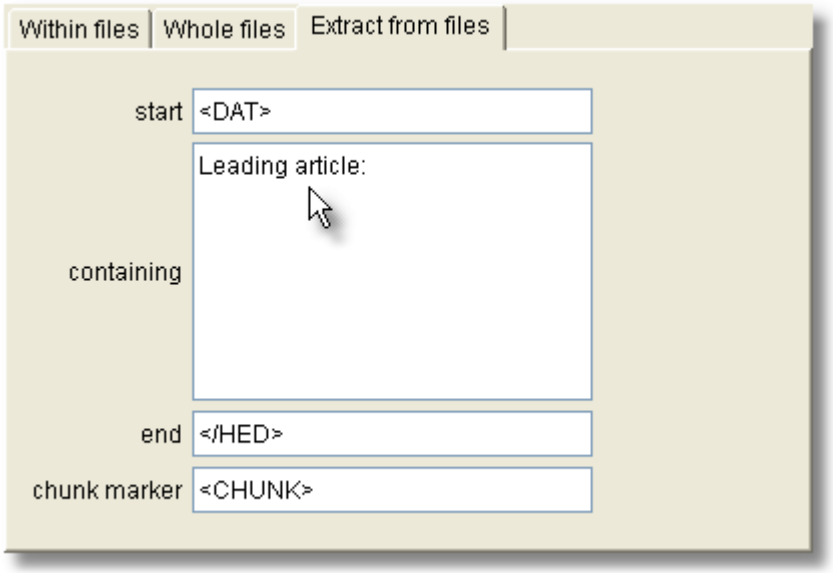
```
c:\temp\shakespeare\poetry\sonnet1.txt
c:\temp\shakespeare\poetry\sonnet2.txt
...
c:\temp\french\victor hugo\miserables.txt
c:\temp\french\poetry\baudelaire\le chat.txt
...
```

In other words, for each file successfully converted or filtered, any same directory structure beyond the starting point (**d:\texts\2007\literature** in the example above) will get appended to the destination.

10.9.3 extracting from files

The point of it...

The idea is to be able to extract something useful from within larger files. In the example below, I wanted to extract the headlines only from some newspaper text. I knew that the header for each text contained **<DAT>** (date of publication mark-up) and that the headline ended **</HED>**, and I wanted only those chunks which contained the phrase **Leading article:**.



The results I got looked like this:

<CHUNK "1"><DAT>05 August 2001</DAT>
<SOU>The Observer</SOU>
<PAG>26</PAG>
<HED>Comment: Leading article: Ealing's lessons: Time for steel from the peacemakers</HED></CHUNK>
<CHUNK "2"><DAT>05 August 2001</DAT>
<SOU>The Observer</SOU>
<PAG>26</PAG>
<HED>Comment: Leading article: The free market can't house us all: Why Government has to intervene</HED></CHUNK>

<CHUNK "3"><DAT>05 August 2001</DAT>
<SOU>The Observer</SOU>
<PAG>26</PAG>
<HED>Comment: Leading article: What a turn-on: Cat's whiskers are the bee's knees</HED></CHUNK>

Settings

containing : **all** non-blank lines in this box will be required. Leave it blank if you have no requirement that the chunk you want to extract contains any given word or phrase.
chunk marker : Leave blank, otherwise each chunk will be marked up as in the example above, if it begins with **<** and ends with **>**. The reason for this marker is to enable subsequent splitting 454.

10.9.4 filtering: move if

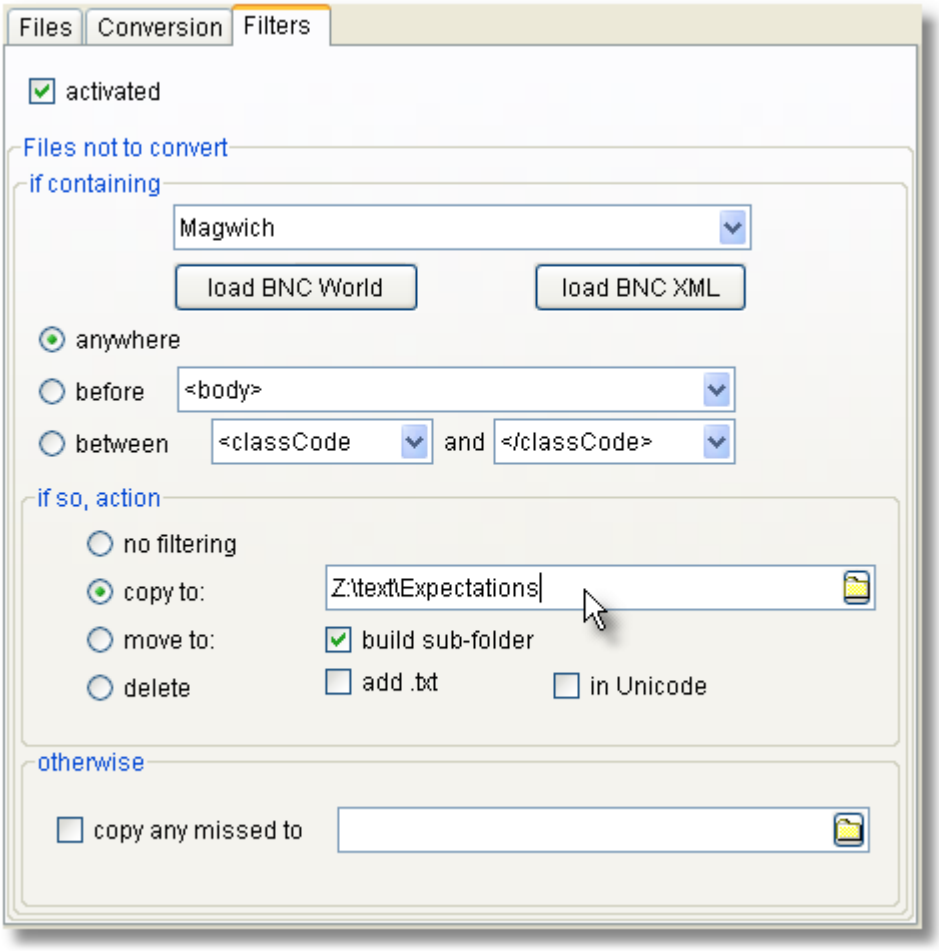
This function allows you to specify a word or phrase, look for it in each file, and if it's found move that file into a new folder.

The point of it …

Suppose you have a whole set of files some of which contain dialogues between Pip and Magwich, others containing references to the Great Wall of China or the anatomy of fleas. You want those with the Pip-Magwich dialogues and you want them to go into a folder called *Expectations*.

How to do it

1. Click on the *Filters* tab (at the top).
2. Now the *Activated* checkbox.



3. Specify a word or phrase the text must contain. This is case sensitive. In this case **`Magwich`** has been specified.
4. Choose whether that word or phrase has to be found
   - anywhere in the text,
   - anywhere before some other word or phrase, or
   - between 2 different words or phrases.
5. Decide what happens if the conditions are met:
   - nothing, i.e. ignore that text file
   - copy to a certain folder, or
   - move to that folder, or
   - delete the file (careful!).

Action options

- You can also decide to build sub-folder(s) based on the word(s) or phrase(s) you chose in #3. (The idea is to get your corpus split up into useful sub-folders whose names mean something to you.) If *build sub-folder* is not checked everything goes into the *copy to* or *move to* folder.
- And you may have the program add **`.txt`** (useful if as with the BNC there are no file extensions) and/or convert it to Unicode.
- You could also have any texts not containing the word **`Magwich`** copied to a specified folder.

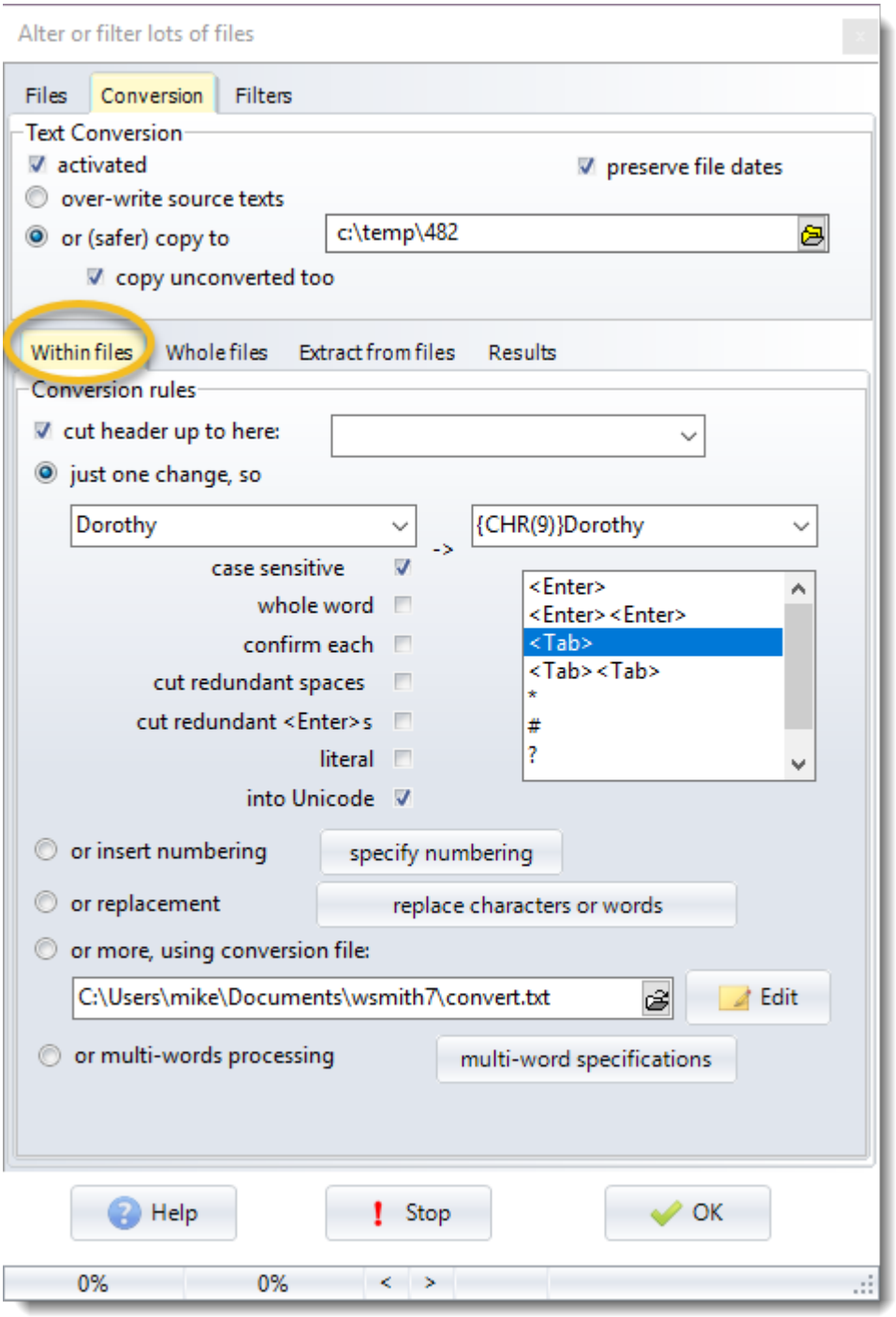
The *load BNC World* and *load BNC XML* buttons are specific to those two editions of the BNC and read text files with similar names which you will find in your **`Documents\wsmith7`** folder.

See also: Text Converter Contents 472.

10.9.5 Convert within the text file

Your choices here are 6:

1. cut out a header



and/or

2. make one change only
3. insert numbering
4. replace some problem characters or words
5. use a script to determine a whole set of changes. There is an sample 486 to see.
6. process any multi-word units 488 in the text using an underscore character, or mark them in your corpus, based on a multi-word unit text file.

If you make one change only you type something into the left box which gets replaced by what is in the right box. In the case above *Dorothy* will get changed to <Tab>+*Dorothy*, that is, the word *Dorothy* will get a tab inserted to its left. The tab was inserted simply by dragging it to the box above it, and when that happened {CHR(9)} appeared automatically being the syntax for a <Tab>. If you know the decimal number for a character you can specify it as {CHR(n)} or simply #n where n represents your number.

With one change only you can see some further options: to force the search-and-replace to be case sensitive, to treat the search as a whole-word search, to cut out any cases where there is more than one space or more than one <Enter> (line-break or paragraph end), to treat the search as literal (see syntax 485) and to force the results into Unicode.

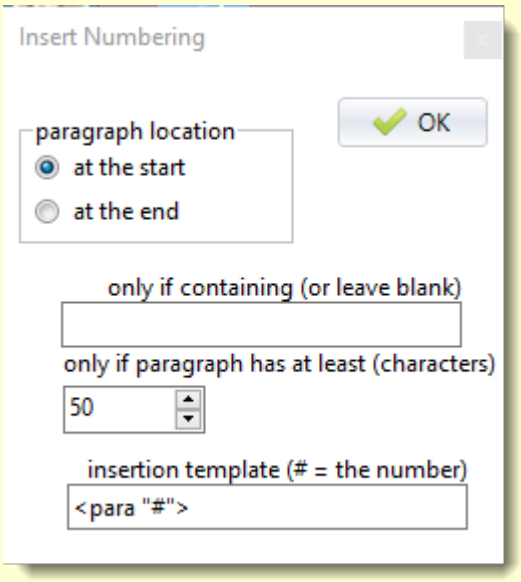
It might be best to check the *confirm each* box too if there's any danger of confusing two different Dorothies with each other.

safer: copy | copy unconverted too

Tip: if copying, check the *copy the unconverted too* box if you want to keep your whole corpus in one place. Any unconverted texts get copied to a sub-folder you you still know which ones were altered.

Insert numbering

This allows you to insert paragraph numbering into your corpus texts. When you click the *specify numbering* button you'll get options like these:



With these choices, for each of your texts, a string like **<para "1">**, **<para "2">** etc. will get inserted at its start if the paragraph has at least 50 characters. The "only if containing" box allows you to specify that numbers only get inserted into paragraphs containing a particular (case-sensitive) string of your choice, such as **Ulan Bator**.
Paragraphs here are identified simply as sequences ending in one <Enter>.

Replace any of these

Allows you to specify a number of characters which you wish to replace, either with a character of your choice or to remove completely if you leave the *with this character* space blank.

See also: curly quotes 491, converting profile texts 407 convert whole file 489, sample conversion file 486, syntax 485, Text Converter Contents 472.

10.9.5.1 conversion file

Prepare your Text Converter conversion file using a plain text editor such as Notepad.
You could use **Documents\wsmith7\convert.txt** as a basis.

If you have accented characters 527 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 486, syntax 485, Text Converter Contents 472.

10.9.5.2 syntax

The syntax for a Conversion File 484 is:

- Only lines beginning / or " are used. Others are ignored completely.
- Every string for conversion is of the form "A" -> "B". That is, the original string, the one you're searching for, enclosed in double quotes, is followed by a space, a hyphen, the > symbol, and the replacement string.
- You can use " (double quotes) and hyphen where you like without any need to substitute them, but for obvious reasons there must not be a sequence like **" -> "** in your search or replace string.

Removing all tags

To remove all tags, choose **"<*>" -> ""** as your search string.

Control Codes

Control codes can be symbolised like this: {CHR(xxx)} where xxx is the number of the code. Examples: **{CHR(13)}** is a carriage-return, **{CHR(10)}** is a line-feed, **{CHR(9)}** is a tab. To represent <*Enter*> which comes at the end of paragraphs and sometimes at the end of each line, you'd type **{CHR(13)}{CHR(10)}** which is carriage-return followed immediately by line-feed.
Use **{CHR(34)}** if you need to refer to double inverted commas. See search-word syntax 246 for more.

Wildcards

The search uses the same mechanism that Concord uses. You may use the same wildcards as in Concord search-word syntax 246. By default the search-and-replace operates on whole words.

Examples:

"book" -> "bk" will replace **book** with **bk** but won't replace **books** or **textbook**

"*book" -> "bk" will replace **book** or **textbook** with **bk** but won't replace **books** or **textbooks**

"book*" -> "bk" will replace **book** or **books** with **bk** but won't replace **textbook** or **textbooks**

To show a character is to be taken literally, put it in quotes (e.g. "*","<"). See below for use of the /L parameter.

Unbounded, case Insensitive, Confirm, redundant Spaces, redundant <Enter>s

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

/U does an unbounded search (ensuring the alteration happens whether there's a word separator 534 on either side or not) (/U "the" finds *the* but also finds *other, then* and *bathe*).

/I does a case insensitive search (/I "restaurant" -> "hotel" replaces *restaurant* with *hotel* and *RESTAURANT* with *HOTEL* and *Restaurant* with *Hotel*, i.e. respecting case as far as possible).

/S cuts out all redundant spaces. That is, it will remove spaces at the beginning of lines, reduce any sequence of two or more spaces to one, and remove and spaces before punctuation. **/S** can be used on a line of its own or in combination with other searches.

/E cuts out all redundant <Enter>s. That is, it will remove empty lines and reduce any sequence of two or more carriage-return+line-feeds (what you get when you press Enter or Return) to one. **/E** can be used on a line of its own or in combination with other searches.

/L means both the search and replace strings are to be taken as literal. (Normally a sequence like **<#*>** would need quotes around each character because **< >** are mark-up signals and **#** and ***** are special wildcard characters, thus **"<""#""*"">"** which is tricky! Put **/L** at the start of the line to avoid this.)

You can combine these, e.g.

```
/IC "the" -> "this"
```

See **Documents\wsmith7 \convert.txt** to see examples in use.

See also: Text Converter Contents 472.

10.9.5.3 sample conversion file

You could copy all or part of this to the clipboard 529 and paste it into notepad.

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

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

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

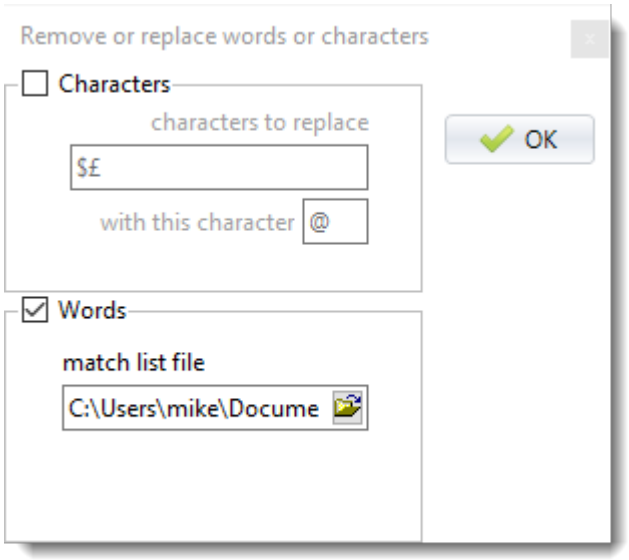
See also: Text Converter Contents 472.

10.9.5.4 remove or replace

This allows you to choose to replace characters or words.

If you choose Characters, you should specify which characters you want to replace and the character to replace them.

Example: the characters setting here shows where a users decides to replace all $ or £ symbols with @.



.

If you choose Words, choose a match list text file. Text Converter will find (case insensitively) each word specified in the match list file and replace it in your source texts with space. That way you could remove all function words from a text.

See also: Match list 82

10.9.5.5 multi-word units

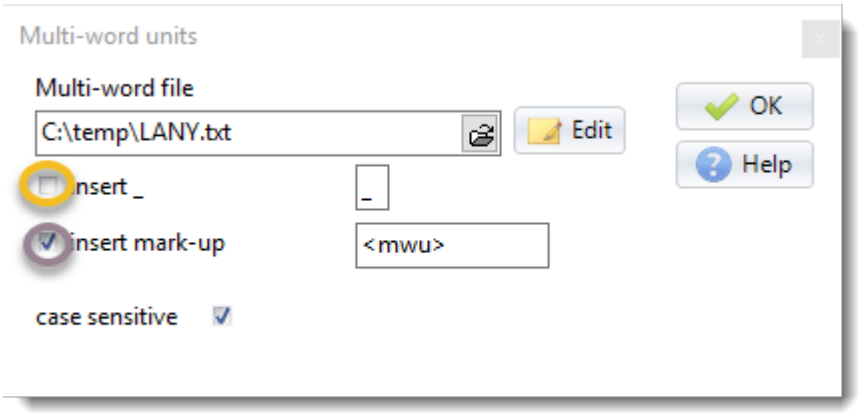
The idea here is to mark up your corpus with clusters or phrases you want treated as single items 306.

You can do that in 2 ways:

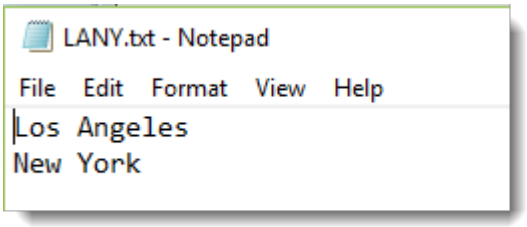
- insert _ so **Los Angeles** becomes **Los_Angeles** and **New York** = **New_York** (yellow ring)

and/or

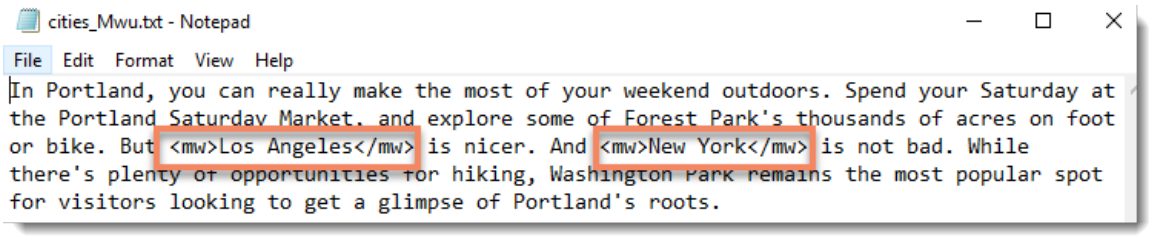
- annotate the text so **Los Angeles** becomes **<mwu>Los Angeles</mwu>**. (grey ring)



For either method you'll need a phrase file which contains the items you're interested in. Mine contained just this:



After processing (using the tag insertion method) my source text looked like this:



with **<mwu>** before and **</mwu>** after each item found.

Case sensitivity

Whether you choose case insensitive or sensitive, the replacement will match the case in your phrase file. If your phrase file has **a lot** then a case insensitive search will also find **A lot** or **a LOT**.

Original: **A lot of people ...**

Conversion: **<mwu>a lot</mwu> of people ...**

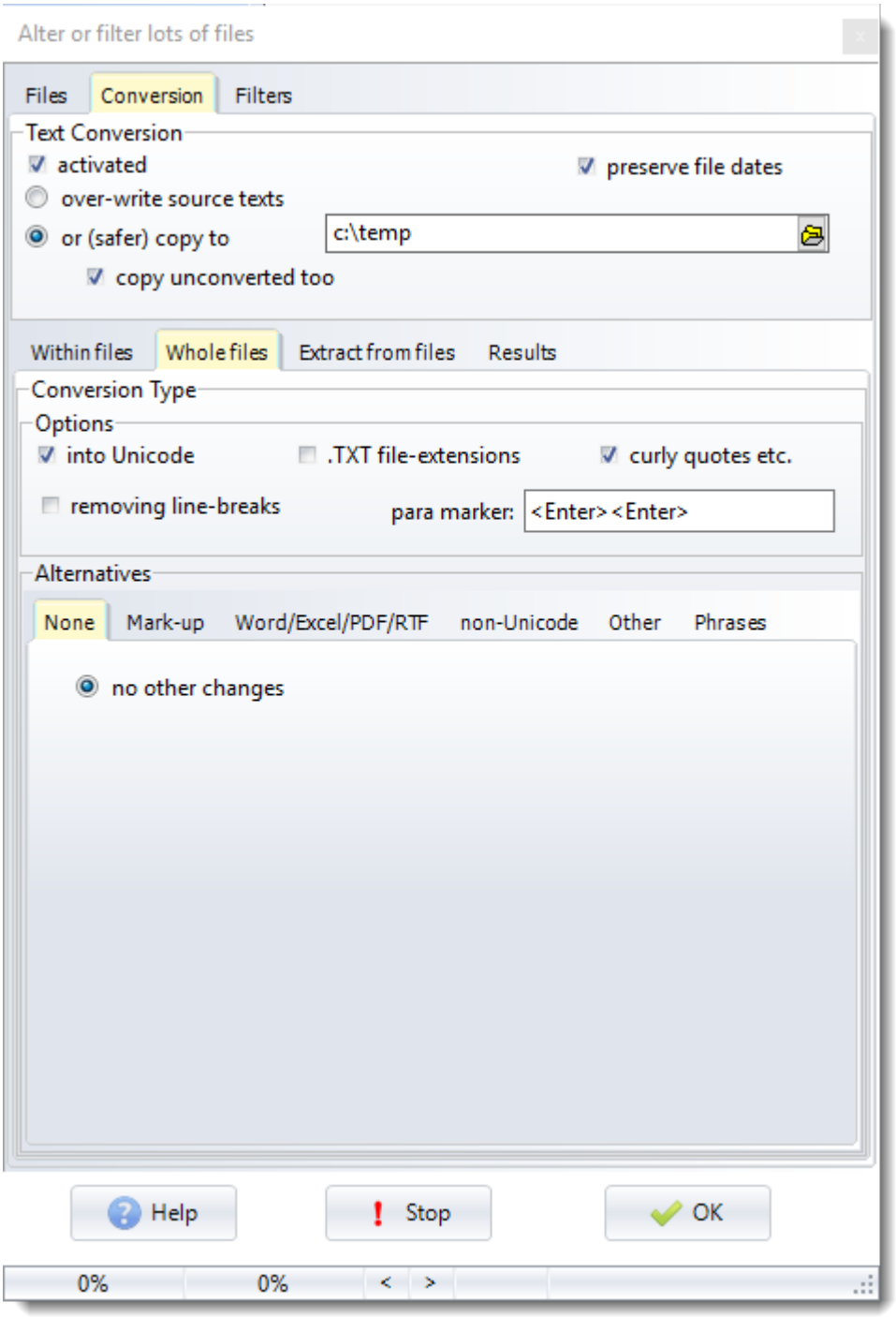
Handling the text now it has been modified

With method 1, you merely need to teach WordSmith that the underscore character is to be accepted 146 as a valid character.

With method 2, you merely have to let WordSmith handle your mark-up 367 to make a word list with clusters in single-word list 306.

10.9.6 Convert format of entire text files

To convert a series of whole text files from one format to another, choose one or more of these options:



Activated : click this to tell the Converter you plan to convert text files.

Preserve file-dates: if checked, text files keep their previous existing date & time after re-saving.

safer: copy | copy unconverted too

Tip: if copying, check the *copy the unconverted too* box if you want to keep your whole corpus in one place. Any unconverted texts get copied to a sub-folder so you still know which ones were altered.

Options

into Unicode (UTF16):

.... this is a better standard than ASCII or ANSI as it allows many more characters to be used, suiting lots of languages. See Flavours of Unicode 538.

TXT file extensions:

... makes the filename end in **.txt** (so that Notepad will open without hassling you; Windows was baffled by the empty filenames of the BNC editions prior to the XML edition). If you choose this you will be asked whether to force **.txt** onto all files regardless, or only ones which have no file extension at all.

curly quotes etc.:

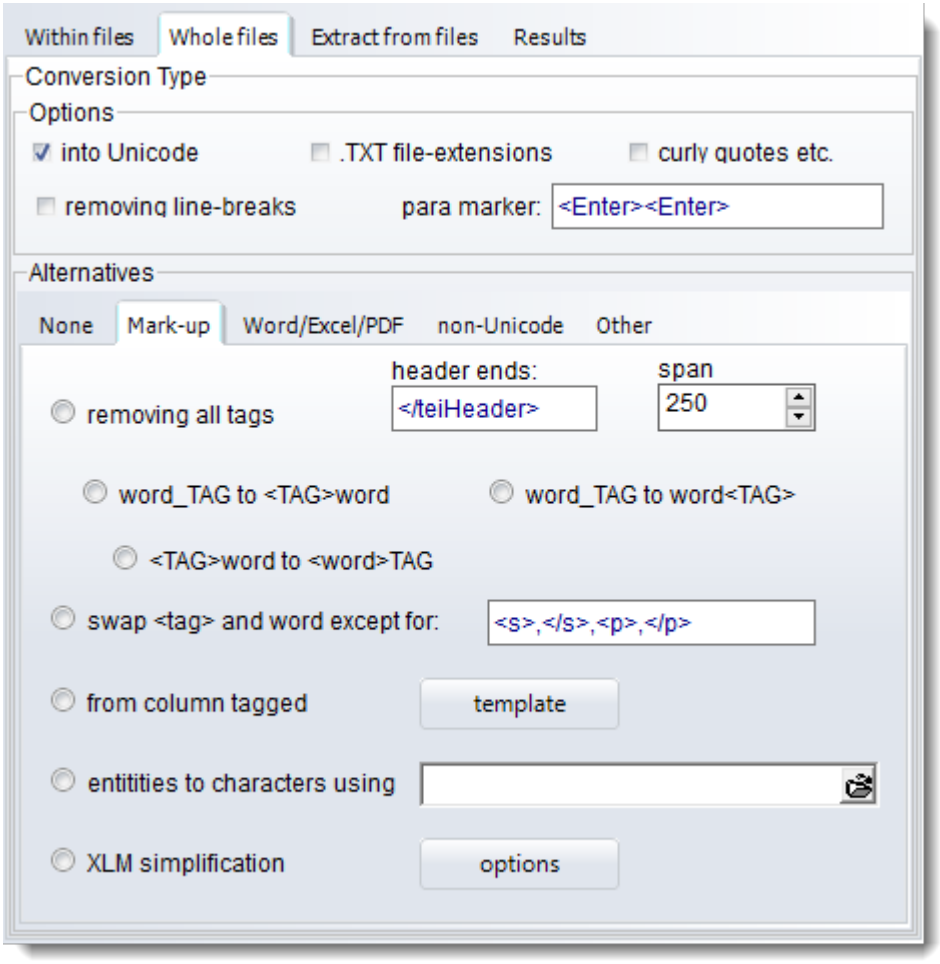
... changes any curly single or double quote marks or apostrophes into straight ones, ellipses into three dots, and dashes into hyphens. (Microsoft's curly apostrophes differ from straight ones.) See also changing a whole set of unwanted characters 407.

removing line-breaks

... replaces every end of line line-break with a space. Preserves any true paragraph breaks, which you must ensure are defined (default = **<Enter><Enter>** -- in other words two line-breaks one after the other with no words between them).

See also: Mark-up 492, Word/Excel/PDF 497, non-Unicode 500, convert within text files 481, change unwanted characters 407, MS Word documents 552, Guide to handling the BNC

10.9.6.1 Mark-up changes



removing all tags

would convert **The<DT><the> TreeTagger<NP><TreeTagger> is<VBZ>**... into **The Treetagger is**. Can plough through a copy of the whole BNC, for example, and make it readable. If you have specified a header string it will cut the header up to that point too. Uses the selected span 159 for looking for the next **>** when it finds a **<**.

word_TAG to <TAG>word

The Helsinki corpus can come tagged like this (COCOA tags)

the_D occasion_N of_P her_PRO$ father's_N$ death_N

and this conversion procedure will change it to

<D>the <N>occasion <P>of <PRO$>her <N$>father's <N>death

You will have to confirm which character such as _ or / divides the word from the existing tags. Note: before it starts, it will clear out any existing <> markup.

word_TAG to word<TAG>

converts text like

It_PP is_VBZ easy_JJ

or Stanford Log-linear POS tagger output like

It/PP is/VBZ easy/JJ

to

It<PP> is<VBZ> easy<JJ>

You will have to confirm which character such as _ or / divides the word from the existing tags. Note: before it starts, it will clear out any existing <> markup.

<TAG>word to TAG<word>

puts the angle brackets round the words instead of the tags.

<II>In <APPGE>my <NN1>case <,>, <PPIS1>I <VM>would <VVI>agree <IW>with <DD1>this

becomes

II<In> APPGE<my> NN1<case> ,<,> PPIS1<I> VM<would> VVI<agree> IW<with> DD1<this>

swap tag and word

converts text like
It<PP> is<VBZ> easy<JJ>
to
<PP>It <VBZ>is <JJ>easy
or vice-versa. In other words swapping the order of tags and words. The procedure effects a swap at each space in the non-tagged text sequence.

Any tags which do not qualify a neighbouring word but for example a whole sentence or a paragraph should not be swapped, so fill in the box to the right with any such tags, using commas to separate them, e.g. **<s>,</s>, , **

from column tagged

The Stuttgart Tree Tagger produces output like this separating 3 aspects of each word with a <tab>:

```
word          pos       lemma
The           DT        the
TreeTagger    NP        TreeTagger
is            VBZ       be
easy          JJ        easy
to            TO        to
use           VB        use
.             SENT      .
```

You will need to supply a template for your conversion.

Template syntax and examples:

1. Any number in the template refers to the data in that column number. (**The** is in column 1 above, **DT** in column 2 of the original.)
2. Only columns mentioned in the template get used in the final output.
3. Separate columns in your template with a / slash.
4. You can add letters and symbols if you like.
5. A space will get added after each line of your original.

Examples:

- the template **1/<3>/<2>** will produce with the cases above **The<the><DT> Treetagger<Treetagger><NP> is<be><VBZ>** etc. (In other words **<>** brackets shown in the template will get added to the text in columns 3 and 2.)
- the template **<POS="2">/1** will produce **<POS="DT">The <POS="NP">Treetagger <POS="VBZ">is** etc. (In other words **<POS="** shown in the template will get added to the beginning of the text in column 2 and **">** after it. That will get followed by column 1 unaltered).

It will present the text as running text, no longer in columns, but with a break every 80 characters.

entities to characters

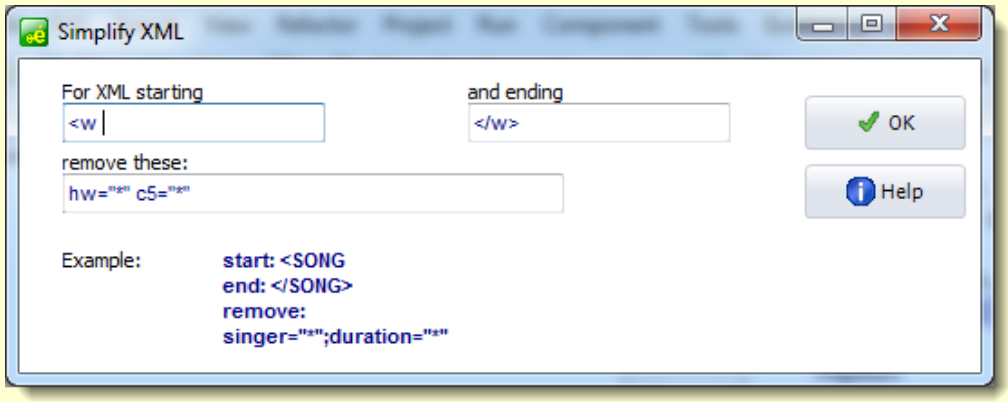
... converts HTML or XML symbols which are hard to read such as **é** to ones like **é**. Specify these in a text file. There is a sample file pre-prepared for you, **html_entities.txt,** in your Documents\wsmith7 folder; look inside and you'll see the syntax.

XML simplification

The idea is to remove any mark-up in XML data which you really do not wish to keep. For example, in the BNC XML edition you might wish to keep only the **pos="*"** mark-up and remove the **c5** and **hw** attributes.

```
<w c5="PRP" hw="in" pos="PREP">In</w>
<w c5="AT0" hw="the" pos="ART">the</w>
<w c5="AJ0" hw="past" pos="ADJ">past</w>
<w c5="NN1" hw="decade" pos="SUBST">decade</w>
<w c5="NP0" hw="amazonia" pos="SUBST">Amazonia</w>
<w c5="VHZ" hw="have" pos="VERB">has</w>
<w c5="VVN" hw="experience" pos="VERB">experienced</w>
<w c5="CRD" hw="one" pos="ADJ">one</w>
<w c5="PRF" hw="of" pos="PREP">of</w>
<w c5="AT0" hw="the" pos="ART">the</w>
<w c5="AJS" hw="big" pos="ADJ">biggest</w>
<w c5="NN1" hw="gold" pos="SUBST">gold</w>
<w c5="VVZ" hw="rush" pos="VERB">rushes</w>
```

To do so, press the Options button and complete for example like this:



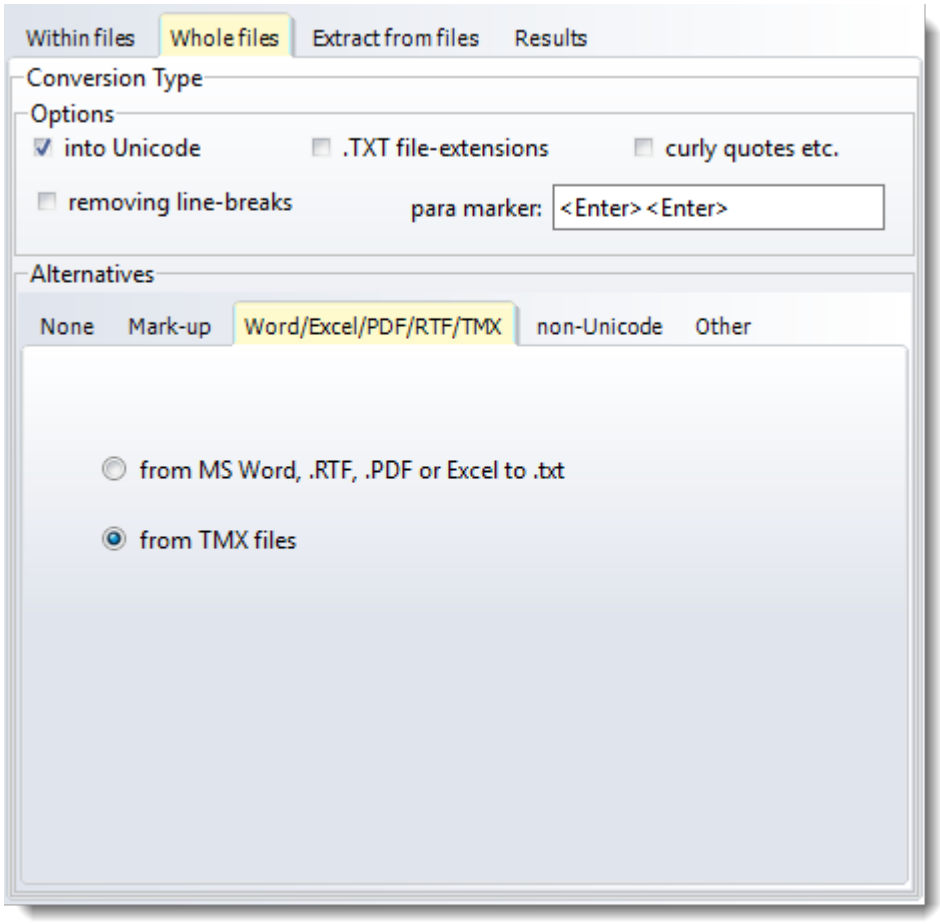
resulting in a saved XML file with a structure like this:

```
<w pos="PREP">In</w>
<w pos="ART">the</w>
<w pos="ADJ">past</w>
<w pos="SUBST">decade</w>
<w pos="SUBST">Amazonia</w>
<w pos="VERB">has</w>
<w pos="VERB">experienced</w>
<w pos="ADJ">one</w>
<w pos="PREP">of</w>
<w pos="ART">the</w>
<w pos="ADJ">biggest</w>
<w pos="SUBST">gold</w>
```

The procedure simply looks for all sections which begin and end with the required strings and delete any sections in between which contain the strings you specify in the *remove these* section. No further account of context is taken. Note that the order of attributes is not important, so we could have specified **c5="*"** first.

See also: Convert Entire Texts 489

10.9.6.2 Word, Excel, PDF, RTF, TMX

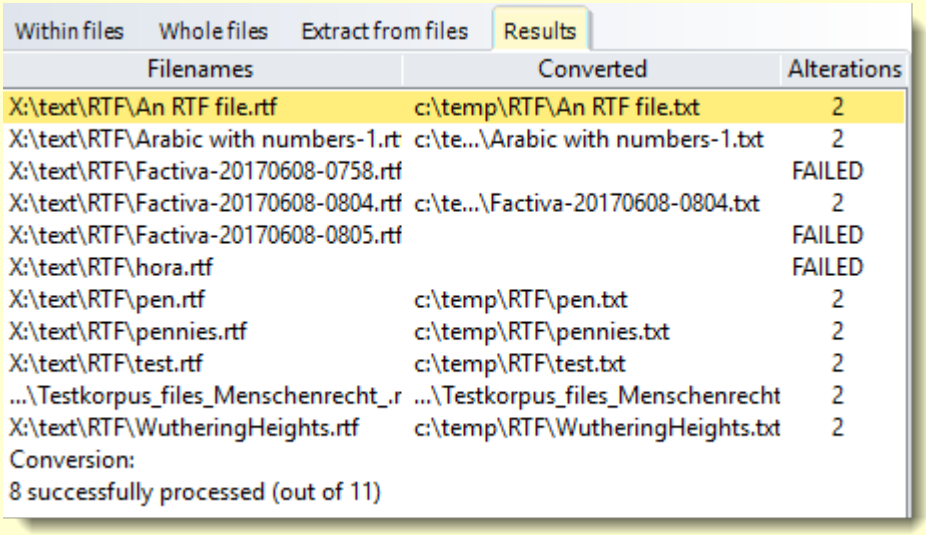


from MS Word, Excel, PDF or RTF to .txt

This is like using "Save as Text" in Word or Excel. Handles .doc, .docx[552] (Office 2007) and **.xls** or **.xlsx** files.

Not guaranteed to work with all .RTF texts but should indicate whether the conversion is or is not successful.

Here is what happens when some conversions fail:



Any which fail you may be able to use the latest version of MS Word to save as text.

from PDF

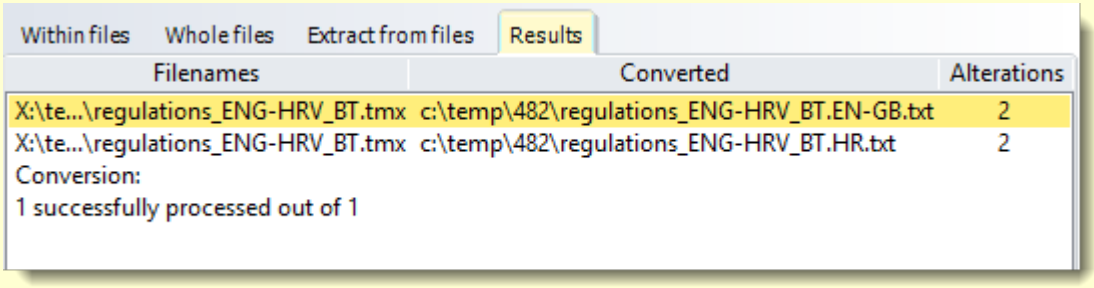
Not guaranteed to work with every .PDF as formats have changed and some are complex.

To convert PDFs to plain text can be extremely tricky even if you own a licensed copy of the Adobe software (Adobe themselves created the PDF format in 1993). That is because PDF is a representation of all the dots, colours, shapes and lines in a document, not a long string of words. It can be very hard with an image of the text, to determine the underlying words and sentences. A second problem is that PDFs can be set with security rights preventing any copying, printing, editing etc.

from TMX

TMX files are translation memory files where one file contains aligned text in more than one language. The idea is to extract the text of the translations in the different languages and save them as separate plain text files.

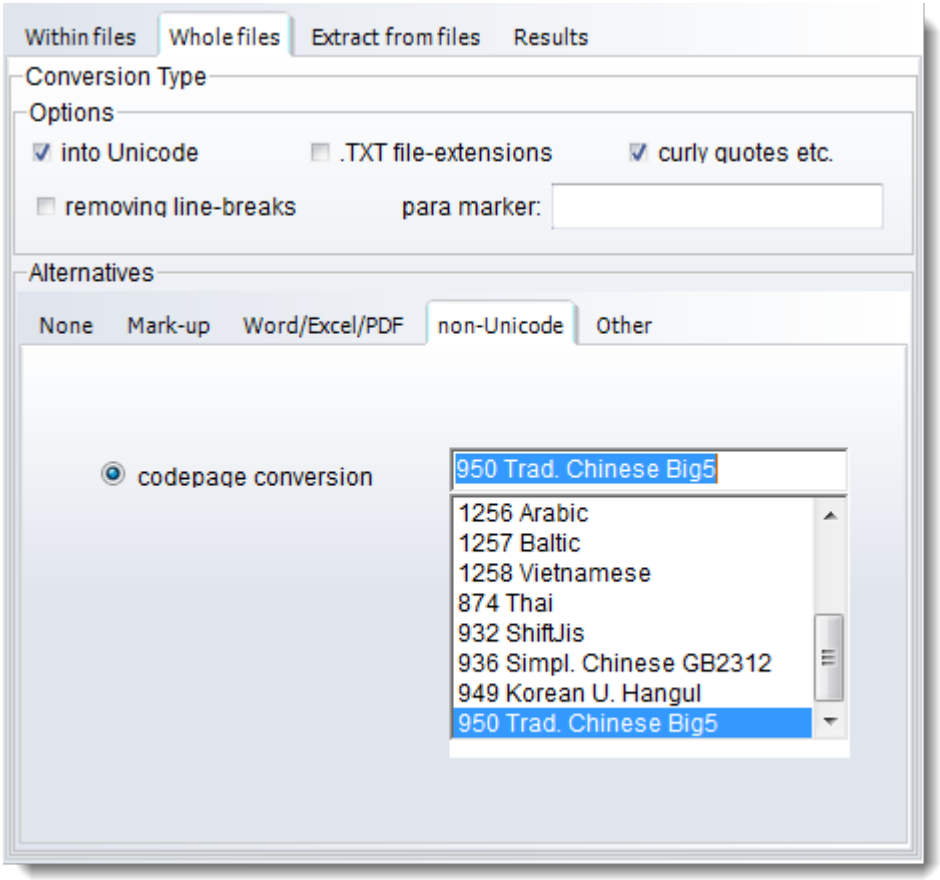
Example



Here one .**TMX** file has had two texts extracted from it, one in English ending **EN-GB.txt** and the other in Croatian ending **HR.txt**. Both are plain text.

See also: Convert Entire Texts 489

10.9.6.3 non-Unicode Text

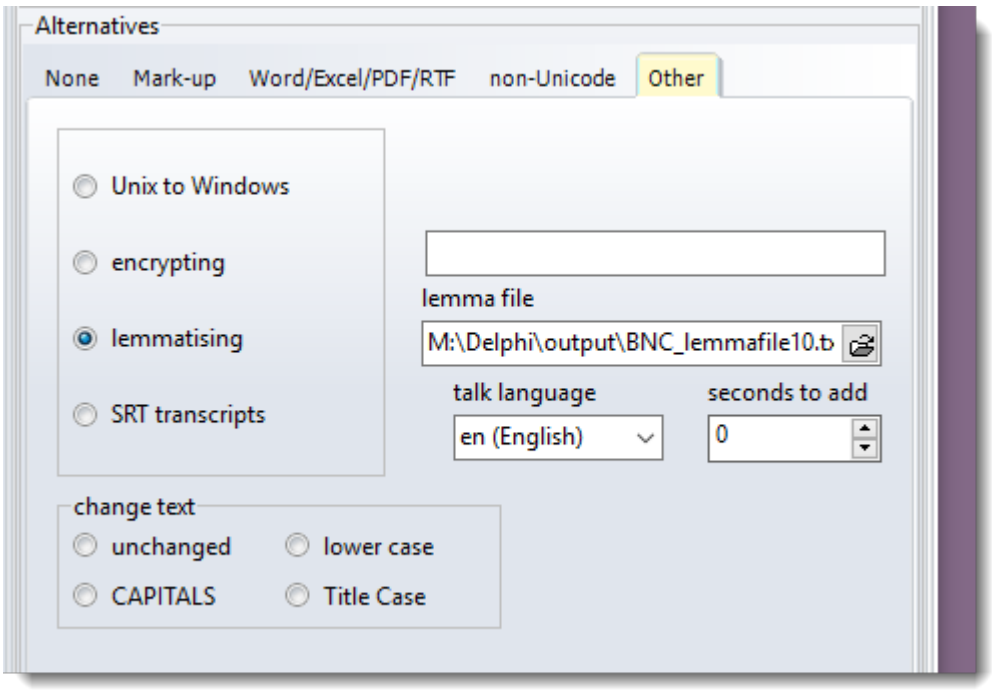


codepage conversion

This allows you to convert 1-byte based formats, for example from Chinese Big5 or GB2312, Japanese ShiftJis, Korean Hangul to Unicode.

See also: Convert Entire Texts 489

10.9.6.4 Other changes



Unix to Windows

Unix-saved texts don't use the same codes for end-of-paragraph as Windows-saved ones. Mac computers use a flavour of Unix.

encrypting using

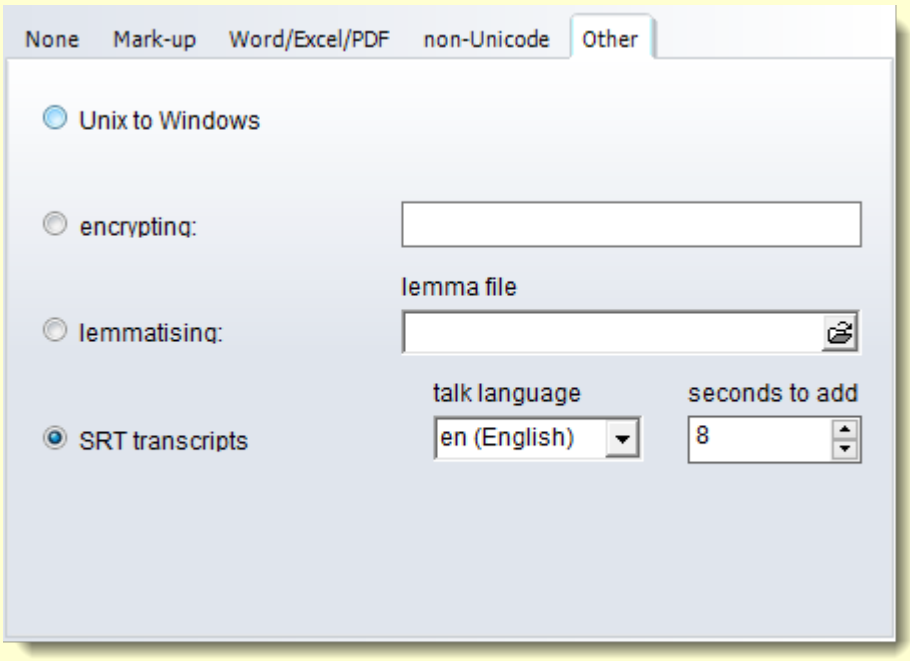
... allows you to encrypt your text files. You supply your own password in the box to the right. When WordSmith processes your text files, e.g. when running a concordance it will restore the text as needed but otherwise the text will be unintelligible. Encrypted files get the file extension **`.WSencrypted`**. For example, if your original is **`wonderful.txt`** the copy will be **`wonderful.WSencrypted`**. Requires the safer *copy to* button above to be selected.

lemmatising using

... converts each file using a lemma file[327]. If for example your source text has "**she was tired**" and your lemma file has **BE -> AM, WAS, WERE, IS, ARE**, then you will get "**she be tired**" in your converted text file. Where your source text has "**Was she tired?**" you'll get "**Be she tired?**"

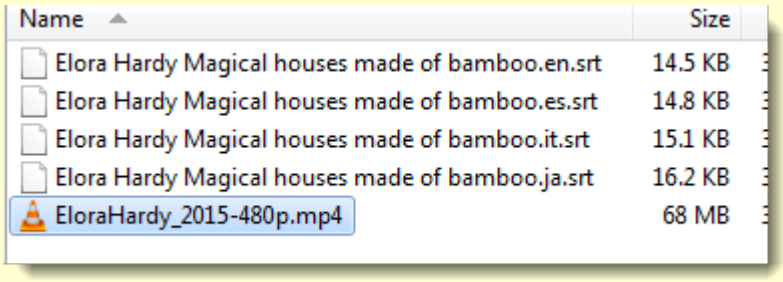
SRT transcripts

... converts SRT files such as those obtained from TED Open Translation Project 263. If using TED files you may need to add about 8 seconds for the standard TED lead-in and you should set the language of the talk too.

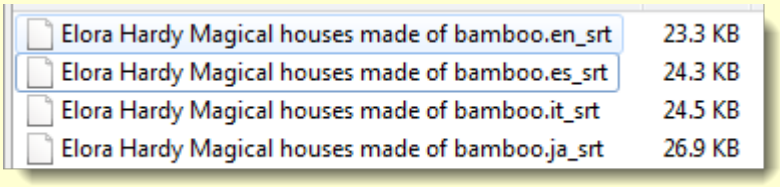


Example

These text files in English (.en), Spanish (.es) Italian (.it) and Japanese (.ja) originally downloaded



got converted thus:



To enable Concord to play the .mp4 file 258 I had to change *EloraHardy_2015-480p.mp4* to the same title *Elora Hardy Magical houses made of bamboo.mp4*. Note the file sizes are bigger (converted into Unicode) and the file-names no longer have two dots. This is so that Concord will find a match between its file-name and the transcripts in these 4 languages.

capitalise, all lower case, Title Case

... converts each text file to capital or lower case letters, or to title case.

See also: Convert Entire Texts 489

10.9.7 Text Converter: converting BNC XML version

The British National Corpus is a valuable resource but has certain problems as it comes straight off the cdrom:

- it is in Unix format
- it has entities like **é** to represent characters like **é**
- its structure is opaque and file-names mean nothing

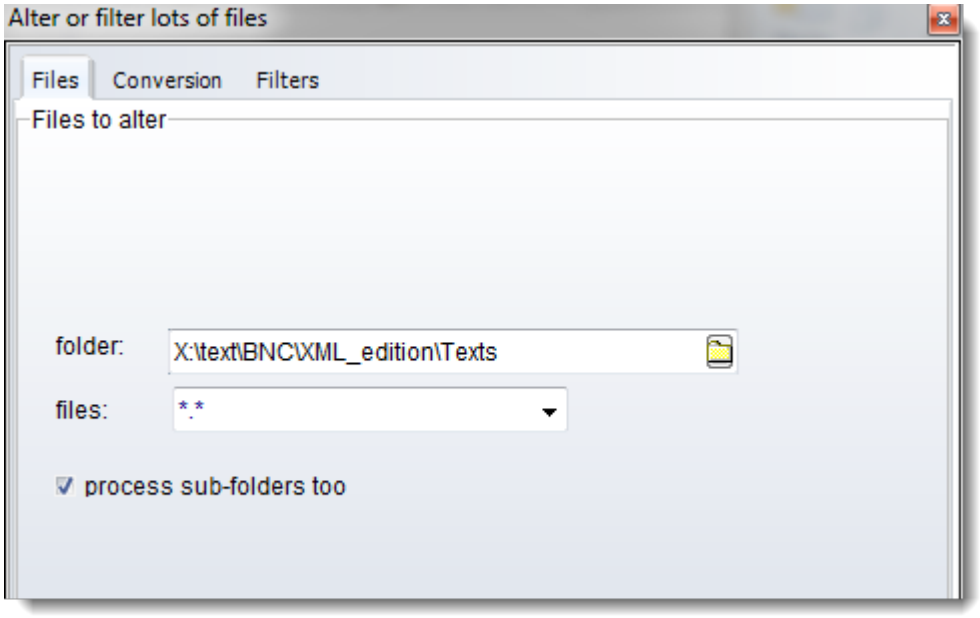
You will find it much easier to use if you

- convert it to Unicode
- filter the files to make a useful structure

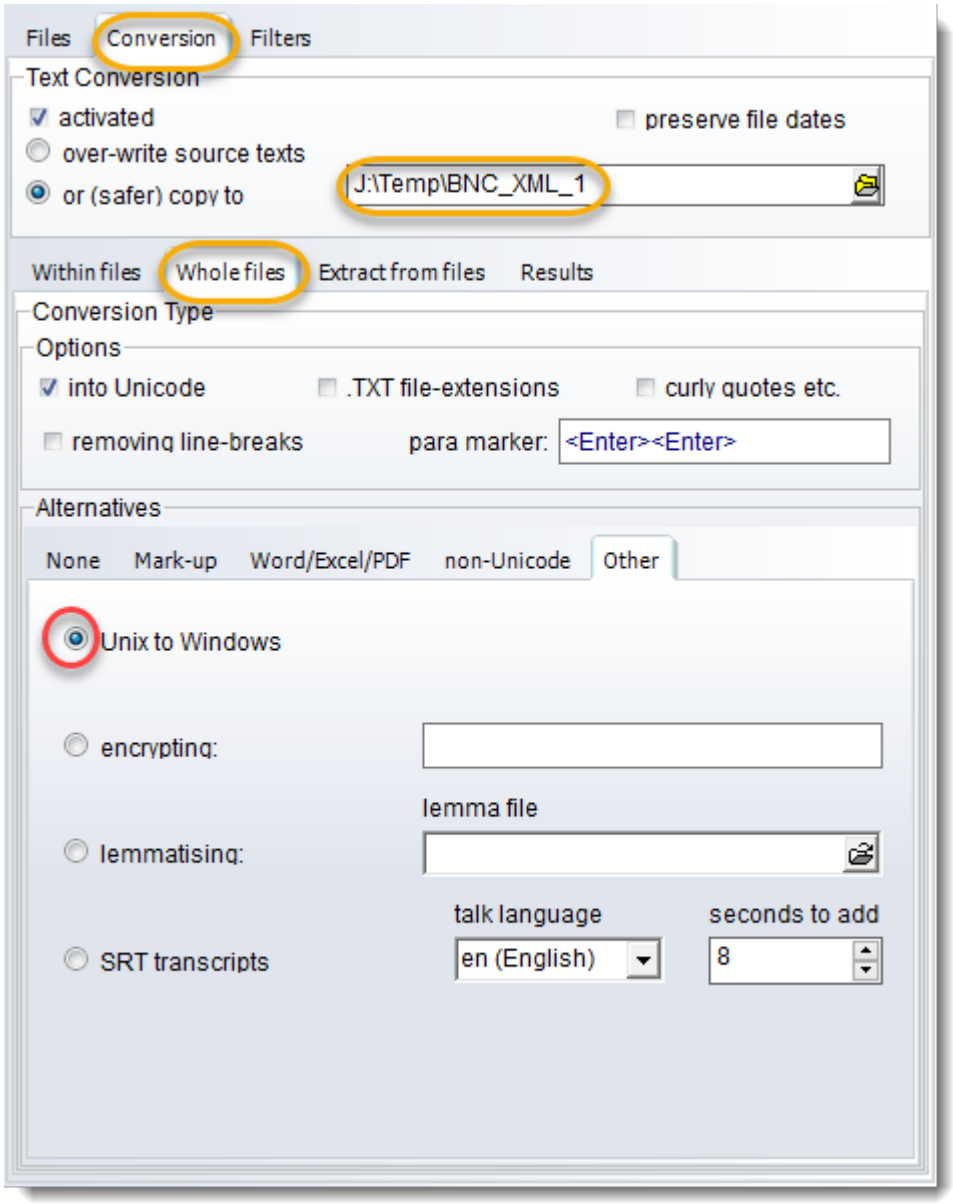
as explained at https://lexically.net/wordsmith/Handling_BNC/index.html

The easiest way to do that is in three stages.

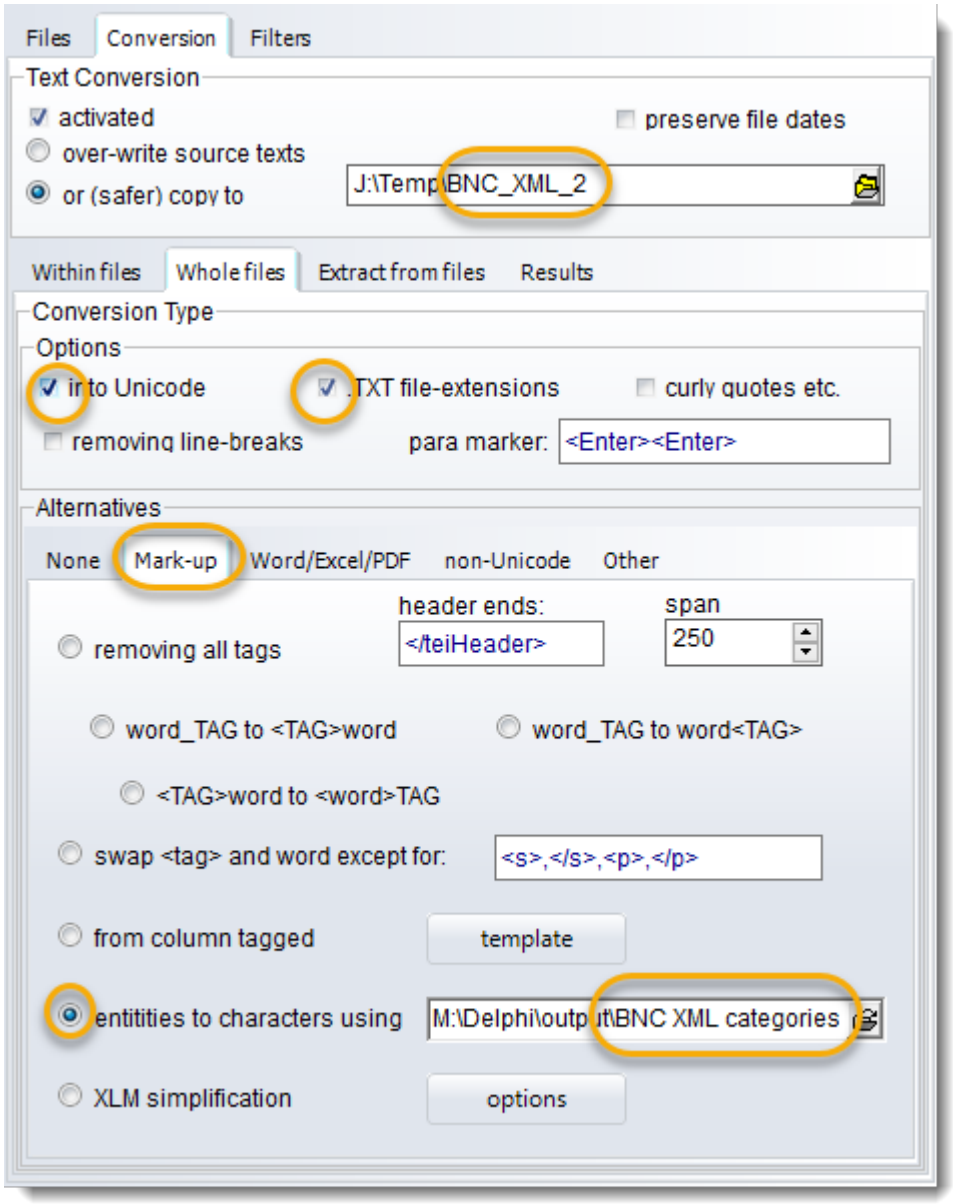
Conversion:



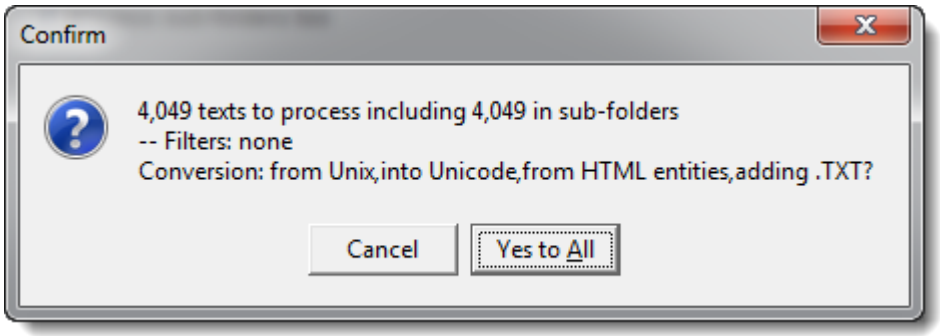
After choosing the texts,



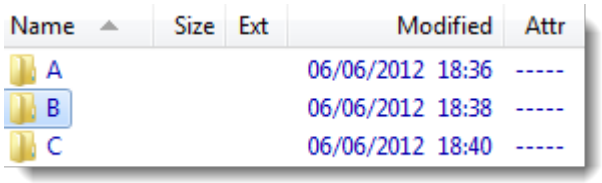
After that, *select the files you have just converted* to Windows format (here at **`J:\temp\BNC_XML_1`**) and do another conversion:

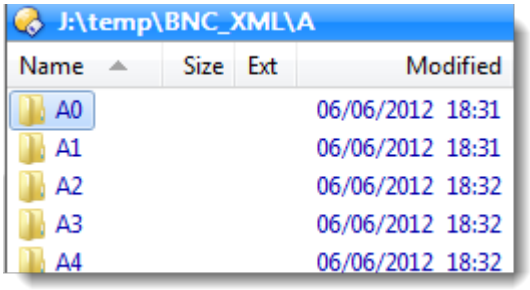


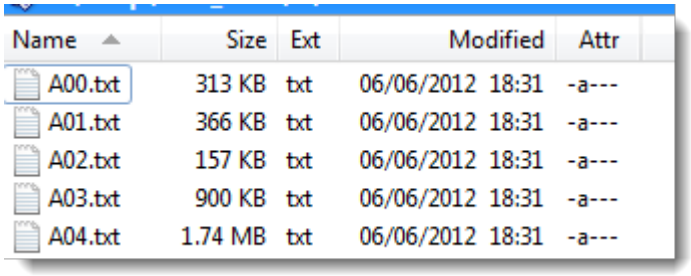
(you will find the BNC XML categories file in your **`Documents\wsmith7`** folder) and when you press OK you'll be asked something like this



After the work is done you will see the BNC texts copied to a similar structure (in our case stemming from **`j:\temp`**)

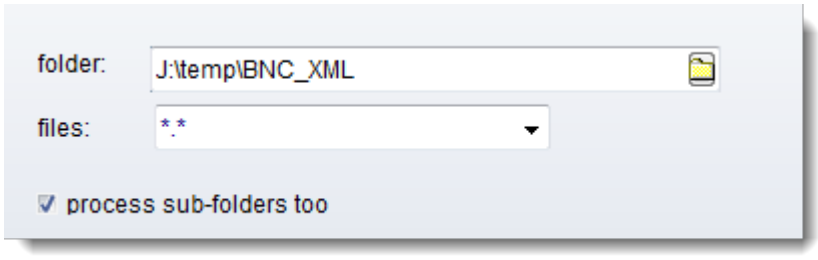




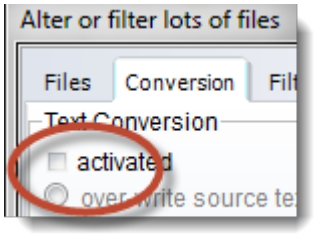


Filter

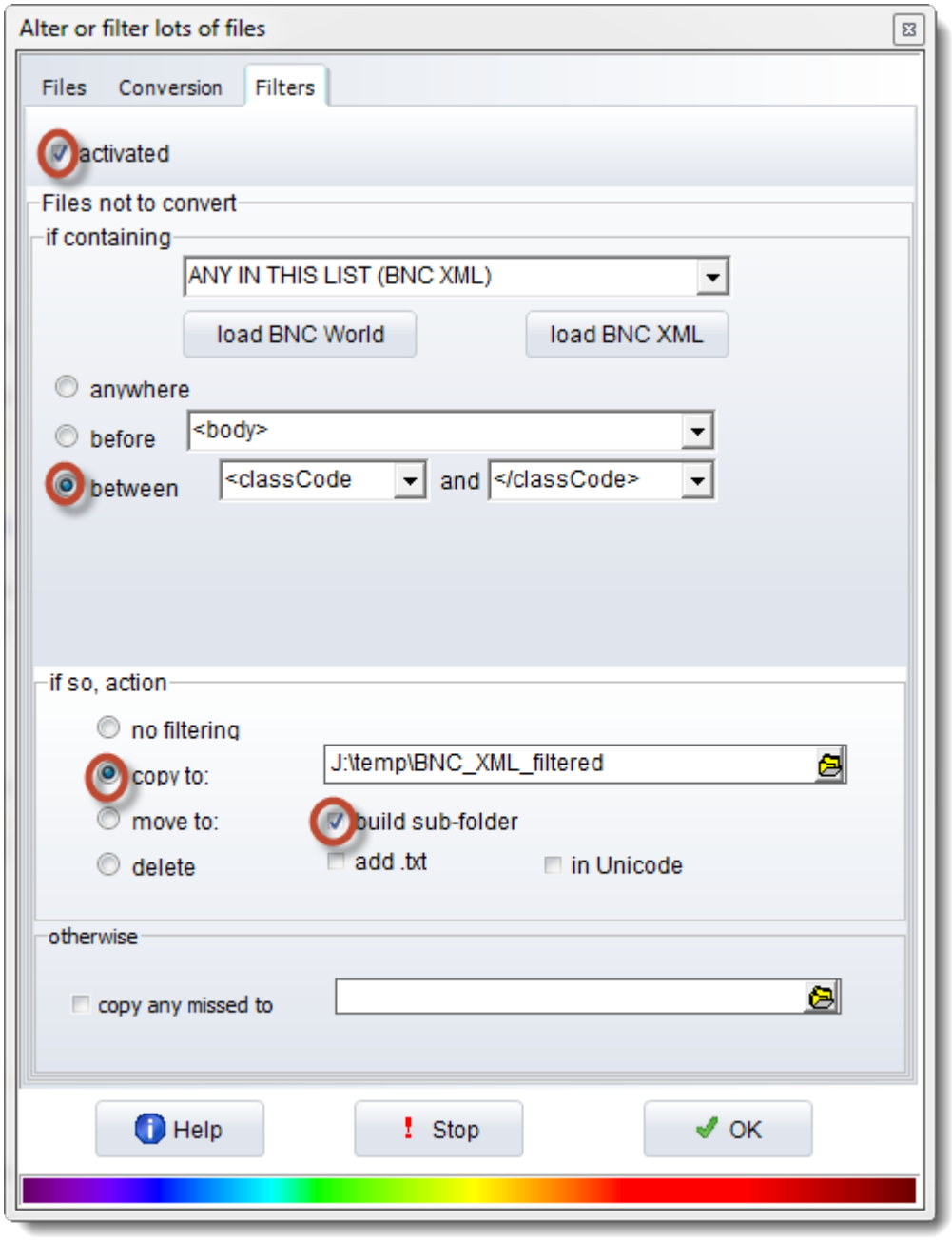
Choose the converted texts in the first window:



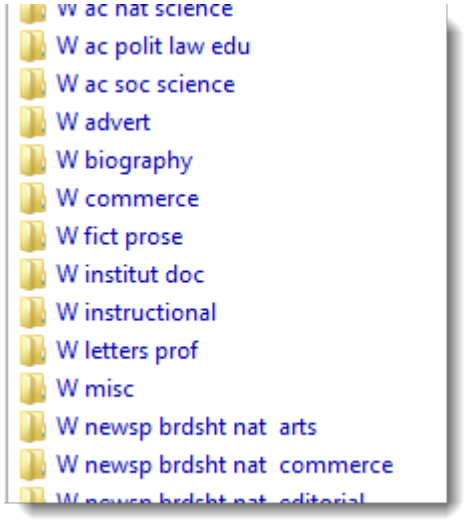
de-activate conversion,



and choose filtering like this:



Eventually you should get folder structures like this:

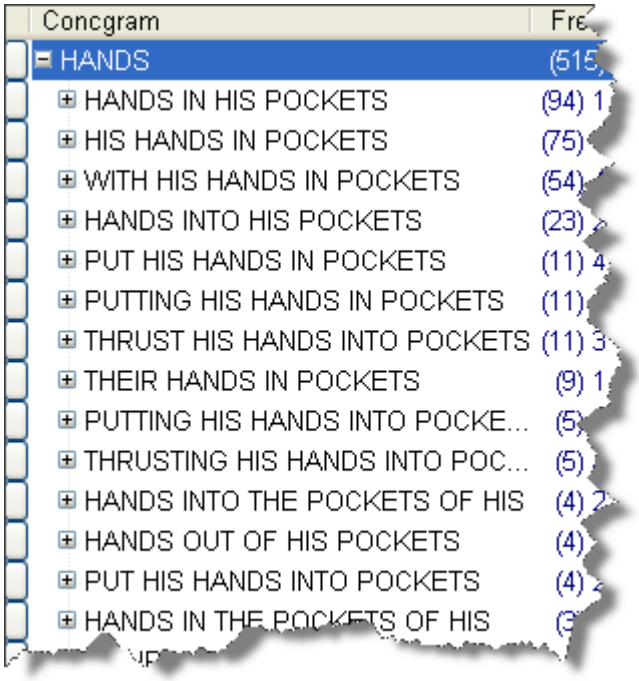


See also: XML simplification 495

10.10 WSConcGram

10.10.1 aims

A program for finding concgrams 511, essentially related pairs, triplets, quadruplets (etc.) of words which are related.



See also : definition of concgram 511, settings 512, running WSConcGram 513, filtering 519, viewing the output 514.

10.10.2 definition of a concgram

For years it has been easy to search for or identify consecutive clusters (n-grams) such as **AT THE END OF, MERRY CHRISTMAS** or **TERM TIME**. It has also been possible to find non-consecutive linkages such as **STRONG** within the horizons 207 of **TEA** by adapting searches to find context words 226. The concgram procedure takes a whole corpus of text and finds all sorts of combinations like the ones above, whether consecutive or not.

Cheng, Greaves & Warren (2006:414) define a concgram like this

> For our purposes, a 'concgram' is all of the permutations of constituency variation and positional variation generated by the association of two or more words. This means that the associated words comprising a particular concgram may be the source of a number of 'collocational patterns' (Sinclair 2004:xxvii). In fact, the hunt for what we term 'concgrams' has a fairly long history dating back to the 1980s (Sinclair 2005, personal communication) when the Cobuild team at the University of Birmingham led by Professor John Sinclair attempted, with limited success, to devise the means to automatically search for non-contiguous sequences of associated words.

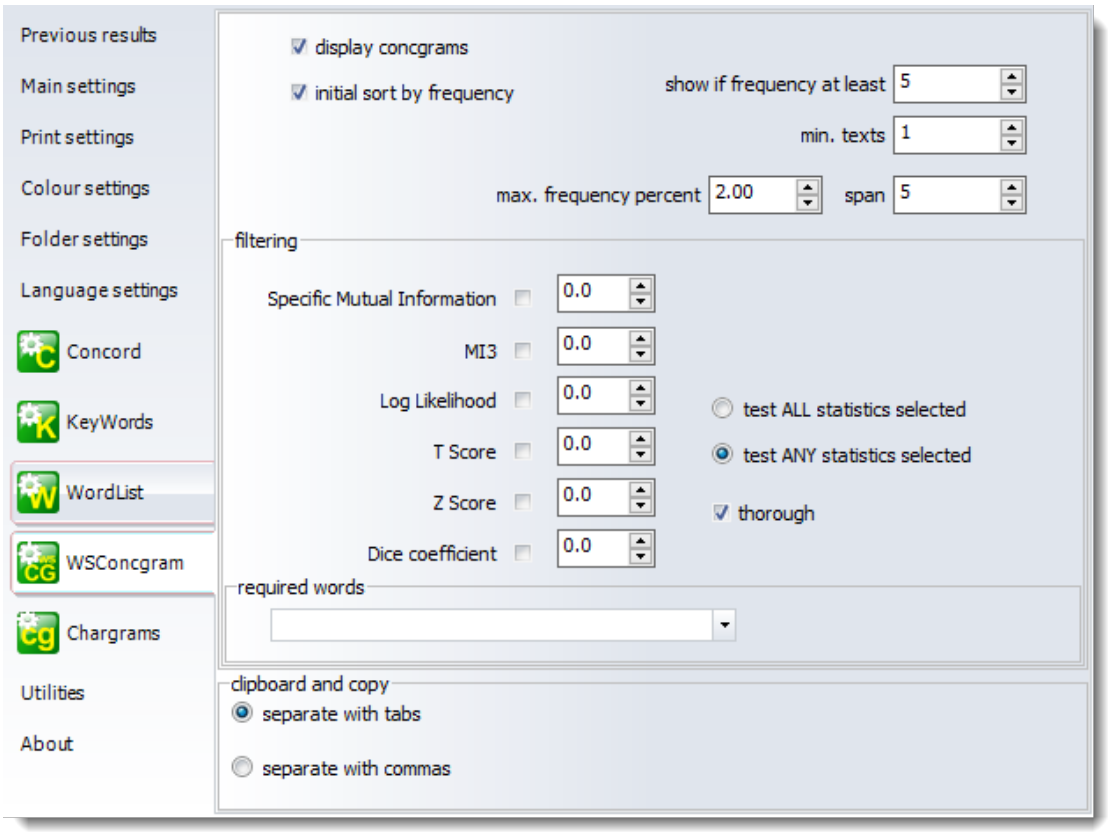
Essentially what they were seeking in developing the ConcGram (©) program was "a search-engine, which on top of the capability to handle constituency variation (i.e. AB, ACB), also handles positional variation (i.e. AB, BA), conducts fully automated searches, and searches for word associations of any size." (2006:413)

WSConcGram is developed in homage to this idea.

See also: bibliography 524, settings 512, running WSConcGram 513, filtering 519, viewing the output 514.

10.10.3 settings

The settings are found in the main Controller.



Computing concgrams

The minimum and maximum settings at the top have to do with the processing ***as*** you carry out the two builds when generating concgrams 513. If you have a lot of data it is best to set a high minimum frequency. The max. frequency percent lets you disregard high-frequency grammar words. A setting of 1.0 means all words whose frequency is 1.0 or higher such as **THE, OF, TO, AND, A** in the word list display 380 would be ignored.

Filter and copy settings

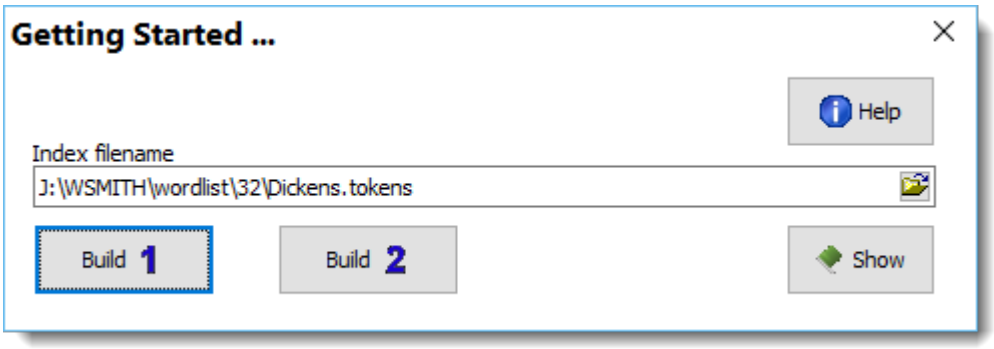
The filtering settings and required words let you require a number of statistical tests ***after*** you have your concgrams computed. These are the same as you choose when computing relationships 344 based on a WordList Index.

Clipboard and copy settings are used when exporting data.

10.10.4 generating concgrams

To start, as usual, choose *File* | *New.*

In the *Getting Started* window, first choose an existing Index, as here where an index based on the works of Dickens has been selected.



To generate the concgrams, the program will then need to build some further files based on the existing index files:

There are two steps simply because there's a lot of work if the original index is large. You can stop after the first stage and resume the next day if you wish. With a modern PC and a source text corpus of only two or three million words, though, it should be possible to generate the files in a matter of a few minutes. A 30 million word corpus with generous settings will take quite a lot longer.

As you see above, 3 large additional files have been generated at the end of the two *Build steps* marked on the buttons in the top window.

Taking too long?

There is a stop button (!) in the status bar at the bottom of your screen.

Press it to get a chance to quit a lengthy process.

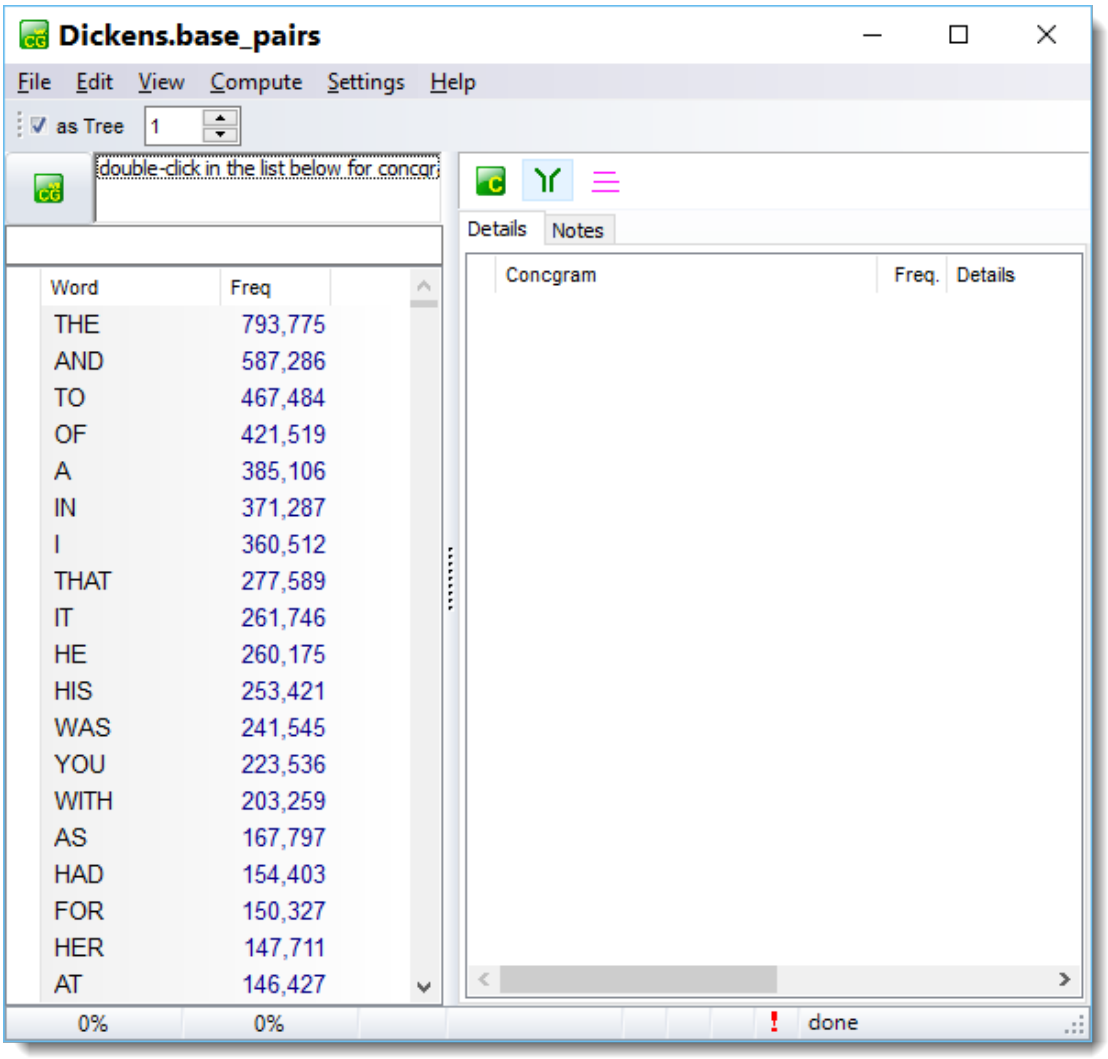
What you get

All items which are found together at least as often as set in the settings (here 512 5 times) will be saved as potential members of each concgram.

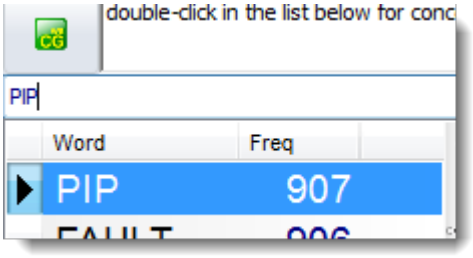
Now, choose Show to view 514 the results. (Or, as usual, right-click the main WSConcgram window and choose *last file*).

10.10.5 viewing concgrams

When you first open a concgram file created by WSConcGram, it will look something like this one

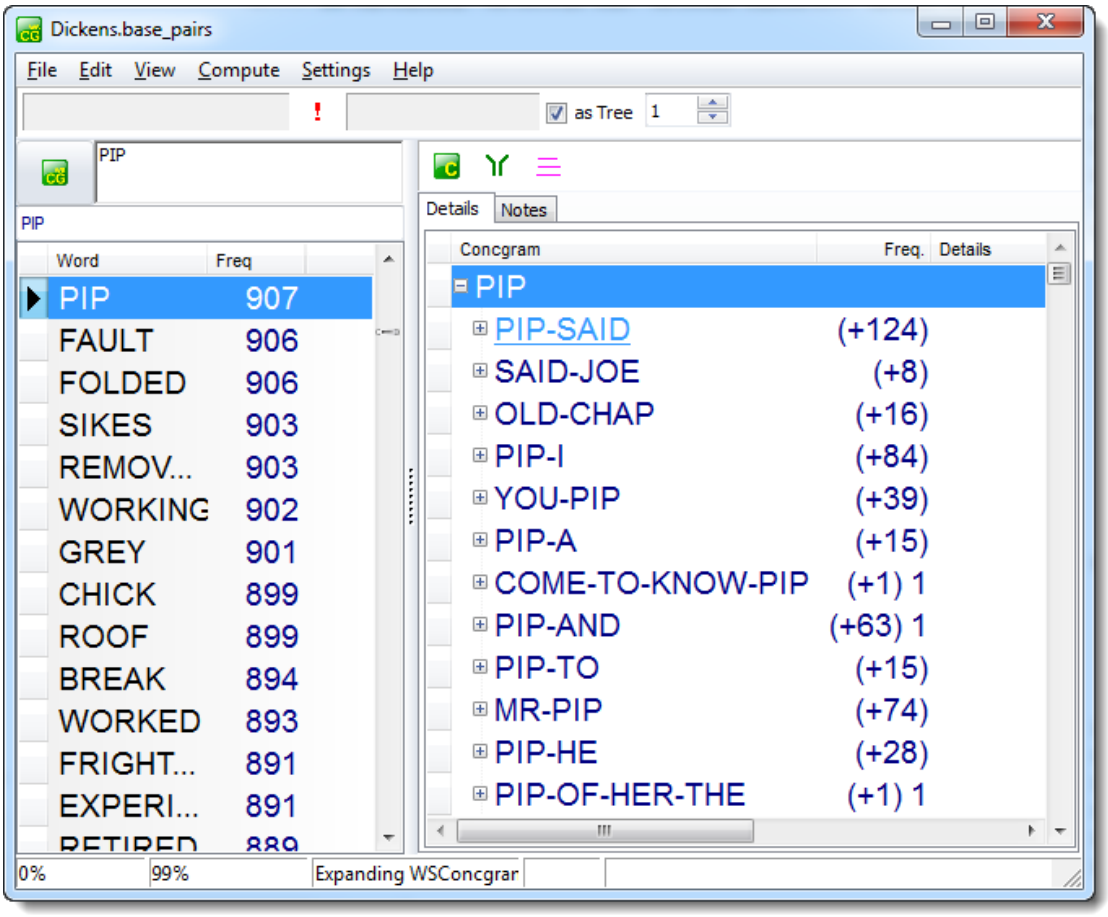


It'll appear (by default) in frequency order as set in the settings 512 but you can sort it by pressing the *Word* and *Freq* headers, and can search for items using the little box above the list.



To get a detailed set of concgrams, double-click an item such as **PIP** (the hero of *Great Expectations*), or drag it to the list-box above. Then press the concgram button beside that.

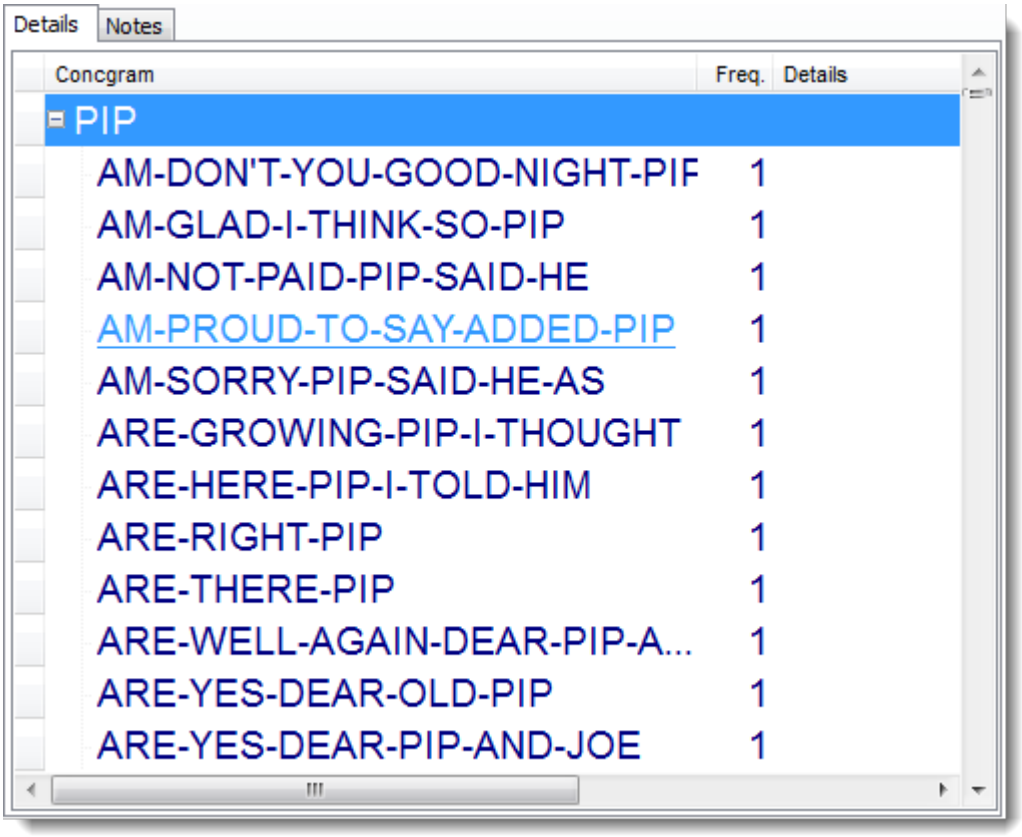
You then get a tree view like this



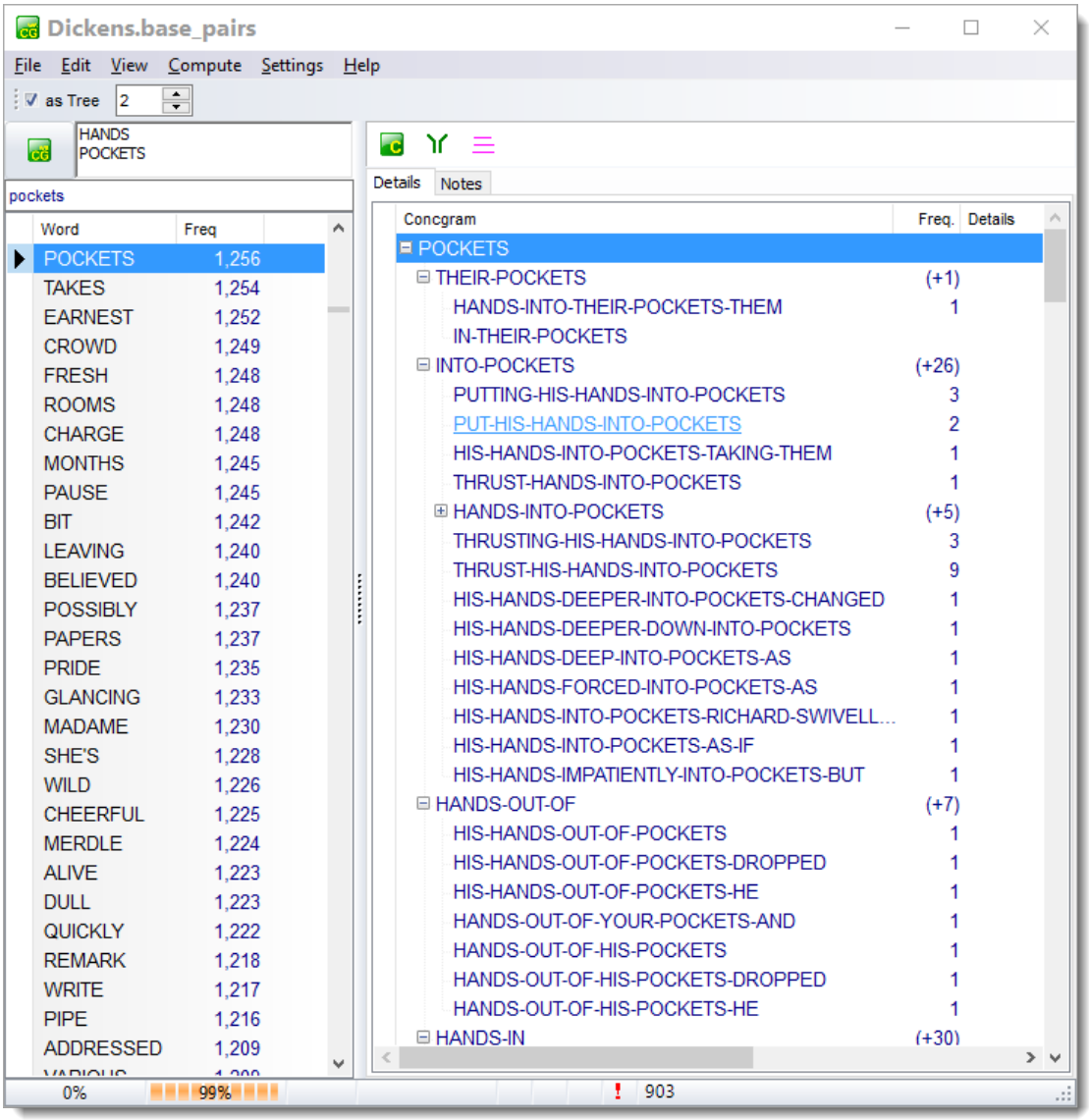
where similar items are grouped. Each branch of the tree shows how many sub-items and how many items of its own it has.

The other controls are used for suspending lengthy processing (!) changing from a tree-view to a list, for concordancing (C), for filtering 519 ( ), clearing filters ( ), and showing more or less of the tree (1).

So if you prefer a plain list, click *as Tree* to view like this:



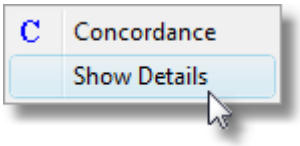
You may if you like select several items like this:



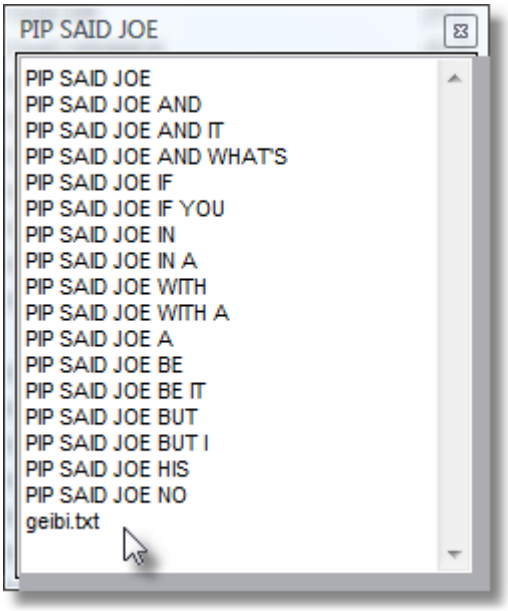
but do note that the concgrams will have to contain *all* of the words selected.

After filtering[519] appropriately and pressing the Concordance button

N	Concordance
1	last. And so GOD bless you, dear old Pip, old chap, GOD bless you!" I had not
2	joy that I knew him. "Which dear old Pip, old chap," said Joe, "you and me
3	me by the old names, the dear "old Pip, old chap," that now were music in
4	I have been ill, Joe," I said. "Dear old Pip, old chap, you're a'most come round,
5	Joe, how smart you are!" "Yes, dear old Pip, old chap." I looked at both of them,
6	the sergeant, staring at Joe. "Halloa, Pip!" said Joe, staring at me. "It was
7	But read the rest, Jo." "The rest, eh, Pip?" said Joe, looking at it with a slow,
8	when you were as little as me?" "Well, Pip," said Joe, taking up the poker, and
9	"There's one thing you may be sure of, Pip," said Joe, after some rumination,
10	see you in your new gen-teel figure too, Pip," said Joe, industriously cutting his
11	it in, and disappeared. "Now, Mr. Pip," said Mr. Jaggers, "attend, if you
12	heart. "Not a particle of evidence, Pip," said Mr. Jaggers, shaking his head
13	through Provis," I replied. "Good day, Pip," said Mr. Jaggers, offering his hand;
14	of to me. "It's a note of two lines, Pip," said Mr. Jaggers, handing it on,
15	bring 'em here, what does it matter?" "Pip," said Mr. Jaggers, laying his hand
16	said that he admitted nothing. "Now, Pip," said Mr. Jaggers, "put this case.
17	some terrible beast. "Look'ee here, Pip. I'm your second father. You're my
18	that's what it was; low. Look'ee here, Pip. Look over it. I ain't a going to be
19	what's due to him. Look'ee here, Pip. I was low; that's what I was; low.
20	who's to gain by it? Still, look'ee here, Pip. If the danger had been fifty times as
21	say, "Ever the best of friends; an't us, Pip? Don't cry, old chap!" When this little
22	last. And so GOD bless you, dear old Pip, old chap, GOD bless you!" I had not
23	I have been ill, Joe," I said. "Dear old Pip, old chap, you're a'most come round,



If you right-click and choose Show Details you'll get to see the details of any section of the tree you have selected:



where you see the various forms and the filename(s) they came from.

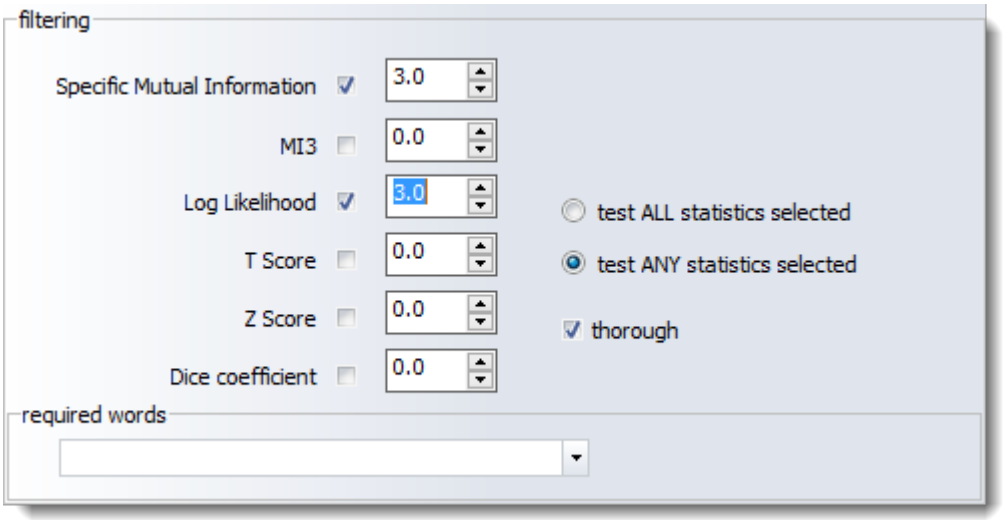
10.10.6 filtering concgrams

The point of it

In order to select which items are "associated", we need some sort of suitable statistical procedures. Before filtering, the members of each concgram are at present merely associated by co-occurring at least a certain minimum number of times, as explained in generating 513 them.

How to do it

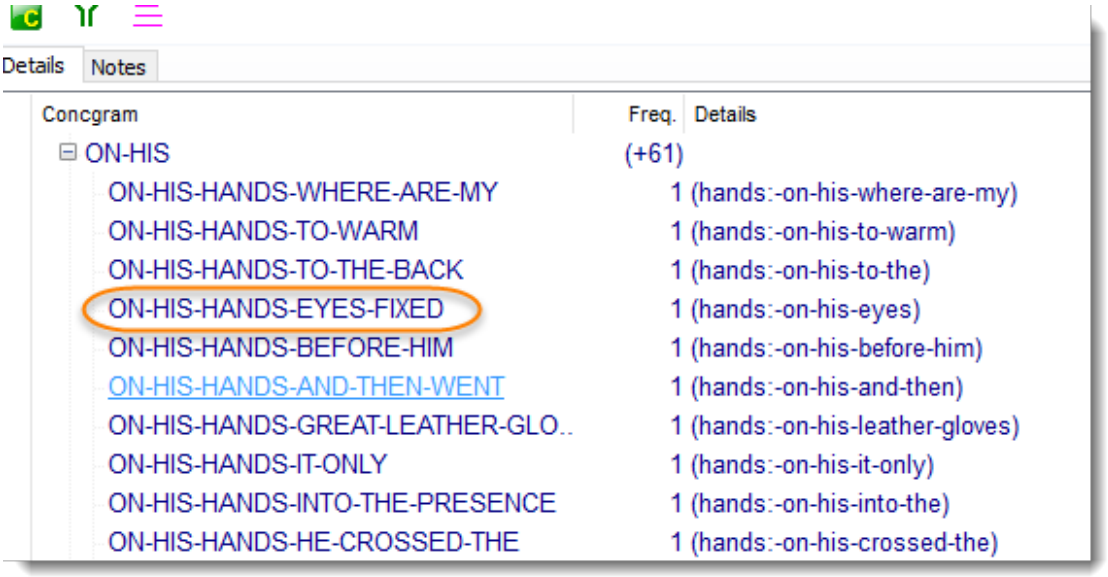
The filtering settings in the Controller allow you to specify, for example, that you want to see only those which are associated with a Specific Mutual Information (MI) score of 3.0 and/or a Log Likelihood score of 3.0.



Ensure the statistics you need are checked and set to suitable thresholds, and decide whether *any one* of the thresholds have to be met (in the case above *either* MI and Log Likelihood would have to score 3.0 at least) or *all* of them (in the case above would require *both* MI at 3.0 or above and Log Likelihood at 3.0 or above). You can also optionally insist on certain words being in your filtered results.

What you see

When you press the filter button ( ), you will see something like this:

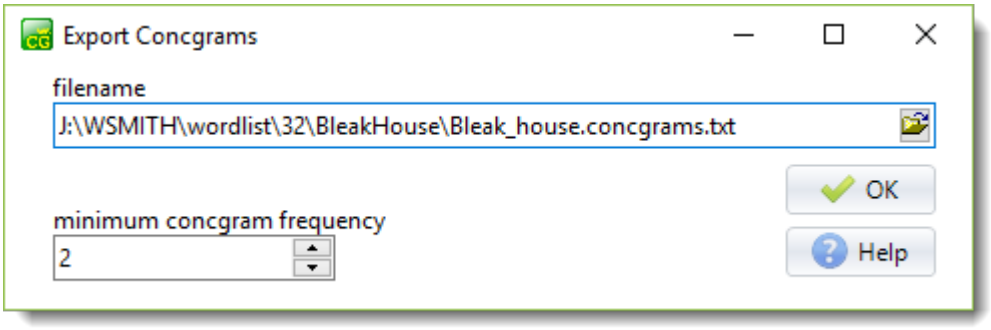


where the items which meet the filter requirements are separated out and selected ready for concordancing; any others are hidden. The *thorough* checkbox setting gives you details of which of the collocates met the requirements.To the right you see that the head-word **HANDS** here relates to **ON, HIS** and **EYES** but not **FIXED** because of the thresholds set when I required both MI and LL of 3.0 for the Dickens **hands** data filtering.

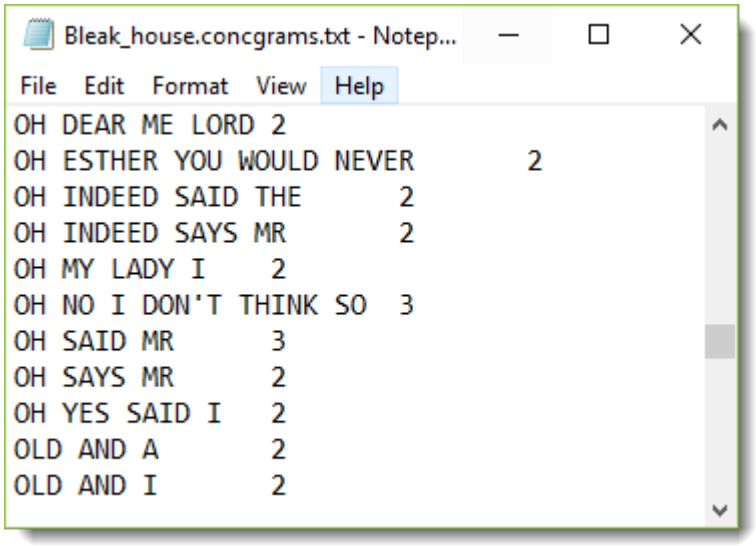
10.10.7 exporting concgrams

With concgram data loaded, you may wish to export it to a plain text file which can be imported into Excel or imported into a WordSmith word-list 363.

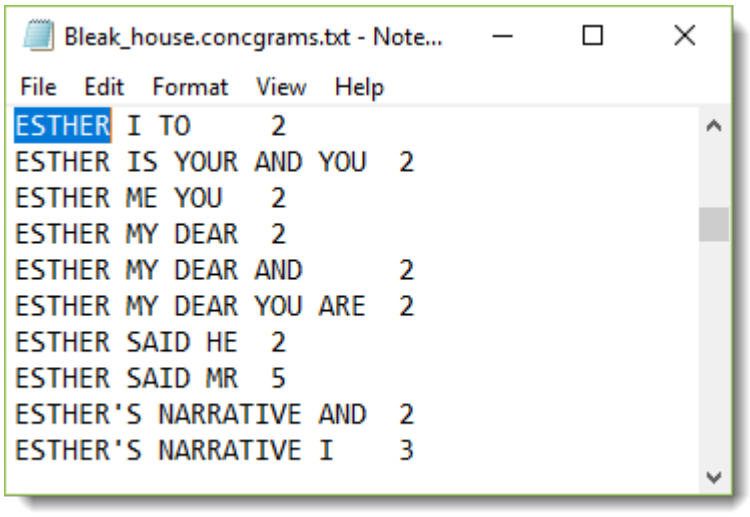
Choose *Compute* | *WordList* and you will be offered choices like these.



The suggested filename is based on your concgram data. You may get a list looking like this:



Section starting **OH, DEAR, OH ESTHER** etc. There is a section starting **ESTHER** of course:



Reference

11 Reference

11.1 acknowledgements

WordSmith Tools has developed over the last few decades. Originally each tool came about because I wanted a tool for a particular job in my work as an Applied Linguist. Early versions were written for DOS, then Windows™ came onto the scene.

One tool, **Concord**, had a slightly different history. It developed out of *MicroConcord* which Tim Johns and I wrote for DOS and which Oxford University Press published in 1993.

The first published version (1996) was written in Borland™ Pascal with the time-critical sections in Assembler. Subsequently the programs were converted to Delphi™ 16-bit; this is a 32-bit only version written in Delphi XE and still using time-critical sections in Assembler.

I am grateful to
- lots of users who have made suggestions and given bug reports, for their feedback on aspects of the suite (including bugs!), and suggestions as to features it should have.
- generations of students and colleagues at the School of English, University of Liverpool, the MA Programme in Applied Linguistics at the Catholic University of São Paulo, colleagues and students at Aston University.
- Audrey Spina, Élodie Guthmann and Julia Hotter for their help with the French & German versions of WS 4.0; Spela Vintar's student for Slovenian; Zhu Yi and others at SFLEP in Shanghai for Mandarin for WS 5.0.
- Robert Jedrzejczyk (http://prog.olsztyn.pl/paslibvlc) for his PasLibVCLPlayer which enables WordSmith to play video.
- Danilo Demarco (http://www.danilodemarco.com/) for his CC link icon.
- Picol (www.picol.org) for icons for text alignment.

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 532 with any further ideas for developing **WordSmith Tools**.

11.2 API

It is possible to run the WordSmith routines from your own programs; for this there's an API published. If you know a programming language, you can call a **`.dll`** which comes with WordSmith and ask it to create a concordance, a word-list or a key words list, which you can then process to suit your own purposes.

Easier, however, is to write a very simple batch script 120 which will run WordSmith unattended.

See also : custom processing 128

11.3 bibliography

Aston, Guy, 1995, Corpora in Language Pedagogy: matching theory and practice, in G. Cook & B. Seidlhofer (eds.) *Principle & Practice in Applied Linguistics: Studies in honour of H.G. Widdowson*, Oxford: Oxford University Press, 257-70.

Aston, Guy & Burnard, Lou, 1998, *The BNC Handbook*, Edinburgh: Edinburgh University Press.

Biber, D., S. Johansson, G. Leech, S. Conrad and E. Finegan, 2000, *Longman Grammar of Spoken and Written English*, Harlow: Addison Wesley Longman.

Clear, Jeremy, 1993, From Firth Principles: computational tools for the study of collocation in M. Baker, G. Francis & E. Tognini-Bonelli (eds.), 1993, *Text and Technology: in honour of John Sinclair*, Philadelphia: John Benjamins, 271-92.

Cheng, Winnie, Chris Greaves & Martin Warren, 2006, From n-gram to skipgram to concgram. *International Journal of Corpus Linguistics*, Vol .11, No. 4, pp. 411-433.

Conrad, S. M., & Biber, D. 2004. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, Vol 20, 2004. available here

Dunning, Ted, 1993. Accurate Methods for the Statistics of Surprise and Coincidence, *Computational Linguistics*, Vol 19, No. 1, pp. 61-74.

Durrant, Philip L, 2008. High Frequency Collocations and Second Language Learning. PhD thesis, University of Nottingham.

Egbert, Jesse & Biber, Doub, 2019. Incorporating text dispersion into key word analyses. *Corpora*, Vol 14 (1), 77-104.

Evert, Stefan, 2007, Corpora and Collocations. mimeo.

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.

Gabrielatos, Costas, 2018. "Keyness Analysis: nature, metrics and techniques". In C. Taylor and A. Marchi (eds.) *Corpus Approaches to Discourse: A critical review*. Chapter 12. Oxford: Routledge.

Gabrielatos, Costas, 2017. Clusters of Keyness. Conference presentation. (PDF based on PPT)

Gries, Stefan Th., 2010. "Useful statistics for corpus linguistics". In Aquilino Sánchez & Moisés Almela (eds.), *A mosaic of corpus linguistics: selected approaches*, 269-291. Frankfurt am Main: Peter Lang.

Gries, Stefan Th., 2013. 50-something years of work on collocations: what is or should be next... *International Journal of Corpus Linguistics*, Vol. 18, No. 1, 137-165.

Katz, Slava, 1996, Distribution of Common Words and Phrases in Text and Language Modelling, *Natural Language Engineering* 2 (1), 15-59

Murison-Bowie, Simon, 1993, *MicroConcord Manual: an introduction to the practices and principles of concordancing in language teaching*, Oxford: Oxford University Press.

Nakamura, Junsaku, 1993, "Statistical Methods and Large Corpora: a new tool for describing text types" in M. Baker, G. Francis & E. Tognini-Bonelli (eds.), 1993, *Text and Technology: in honour of John Sinclair*, Philadelphia: John Benjamins, 293-312.

Oakes, Michael P. 1998, *Statistics for Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Scott, Mike, 1997, "PC Analysis of Key Words - and Key Key Words", *System*, Vol. 25, No. 2, pp. 233-45.

Scott, Mike & Chris Tribble, 2006, *Textual Patterns: keyword and corpus analysis in language education*, Amsterdam: Benjamins.

Sinclair, John M, 1991, *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Stubbs, Michael, 1986, "Lexical Density: A Technique and Some Findings", in M. Coulthard (ed.) Talking About Text: Studies presented to David Brazil on his retirement, *Discourse Analysis Monograph no. 13*, Birmingham: English Language Research, Univ. of Birmingham, 27-42.

Stubbs, Michael, 1995, "Corpus Evidence for Norms of Lexical Collocation", in G. Cook & B. Seidlhofer (eds.) *Principle & Practice in Applied Linguistics: Studies in honour of H.G. Widdowson*, Oxford: Oxford University Press, 245-56.

Tuldava, J. 1995, *Methods in Quantitative Linguistics*, Trier: WVT Wissenschaftlicher Verlag Trier.

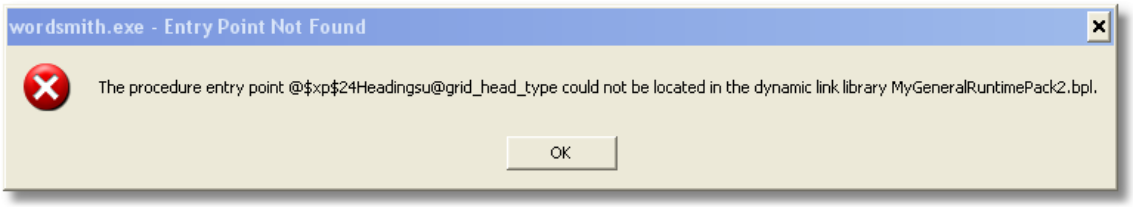
Youlmans, Gilbert, 1991, "A New Tool for Discourse Analysis: the vocabulary-management profile", *Language*, V. 67, No. 4, pp. 763-89.

UCREL's log likelihood information

11.4 bugs

All computer programs contain bugs. You may have seen a "General Protection Fault" message when using big expensive drawing or word-processing packages.

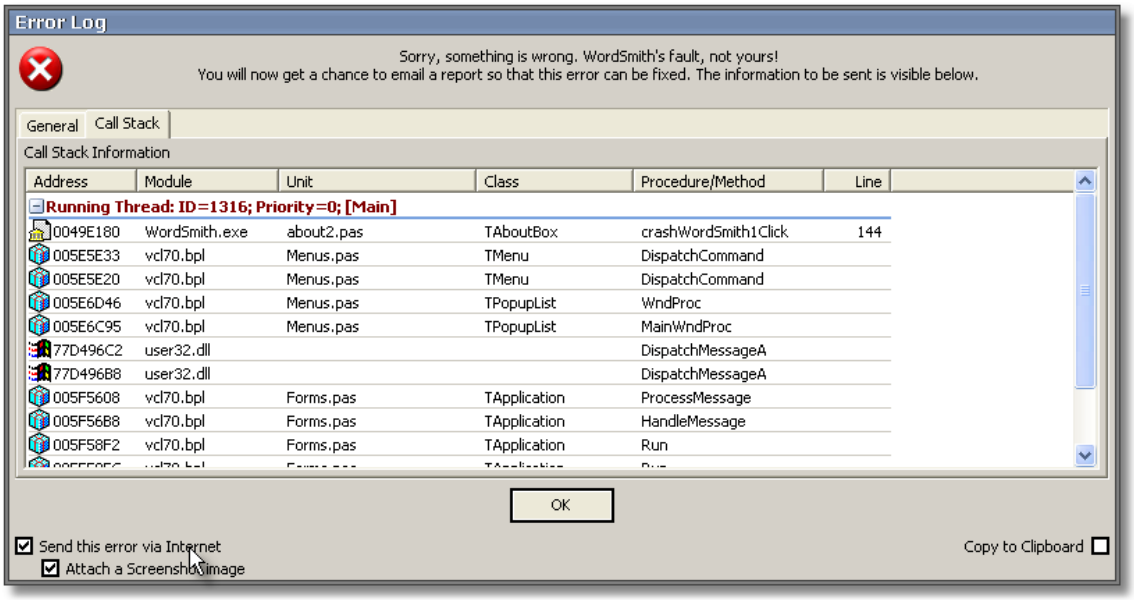
If you see something like this,



then you have an incompatibility between sections of WordSmith. You have probably downloaded a fresh version of some parts of WordSmith but not all, and the various sub-programs are in conflict... The solution is a fresh download.
https://lexically.net/wordsmith/version6/faqs/updating_or_reinstalling.htm explains.

Otherwise you should get a report popping up, giving "General" information about your PC and "Details" about the fault. This information will help me to fix the problem and will be saved in a small text file called **`wordsmith.elf, concord.elf, wordlist.elf`**, etc. When you quit the program, you will be offered a chance to email this to me.

The first thing you'll see when one of these happens is something like this:



You may have to quit when you have pressed OK, or WordSmith may be able to cope despite the problem.
Usually the offending program will be able to cope despite the bug or you can go straight back into it without even needing to quit the main WordSmith Tools Controller 5, retrieve your saved results 91 from disk, and resume. If that doesn't work, try quitting WordSmith Tools overall, or quit Windows and then start it up again.

When you press OK, your email program should have a message with a couple of attachments to send to me.

The email message will only get sent when you press Send in your email program. It is only sent to me and I will not pass it on to anyone else. Read it first if you are worried about revealing your innermost secrets ... it will tell me the operating system, the amount of RAM and hard disk space, the version of WordSmith, and some technical details of routines which it was going through when the crash occurred.

error messages 569

These warn you about problems which occur as the program works, e.g. if there's no room left on your disk, or you type in an impossible file name or a number containing a comma.

See also: logging 120, troubleshooting 563.

11.5 change language

See Controller | Language 79.

11.6 Character Sets

11.6.1 overview

You need "plain text" in WordSmith. Not Microsoft Word .doc 552 files -- which contain text and a whole lot of other things too that you cannot normally see.

If you are processing English only, your texts can be in ASCII, ANSI or Unicode; WordSmith handles both formats. You can also process HTML or XML 543.

Text encoding

To handle a text in a computer, programs need to know how the text is encoded. In its processing, the software sees only a long string of numbers, and these have to match up with what you and I can recognise as "characters". For many languages like English with a restricted alphabet, encoding can be managed with only 1 "byte" per character. On the other hand a language like Chinese, which draws upon a very large array of characters, cannot easily be fitted to a 1-byte system. Hence the creation of other "multi-byte" systems. Obviously if a text in English is encoded in a multi-byte way, it will make a bigger file than one encoded with 1 byte per character, and this is slightly wasteful of disk and memory space. So, at the time of writing, 1-byte character sets are still in very widespread use. UTF-8 is a name for a multi-byte method, widely used for Chinese, etc.

In practice, your texts are likely to be encoded in a Windows 1-byte system, older texts in a DOS 1-byte system, and newer ones, especially in Chinese, Japanese, Greek, in Unicode. What matters most to you is what each character looks like, but WordSmith cannot possibly sort words correctly, or even recognise where a word begins and ends, if the encoding is not correct. WordSmith has to know (or try to find out) which system your texts are encoded in. It can perform certain tests in the background. But as it doesn't actually understand the words it sees, it is much safer for you to convert to Unicode, especially if you process texts in German, Spanish, Russian, Greek, Polish, Japanese, Farsi, Arabic etc.

Three main kinds of character set, each with its own flavours, are Windows, DOS, and Unicode.

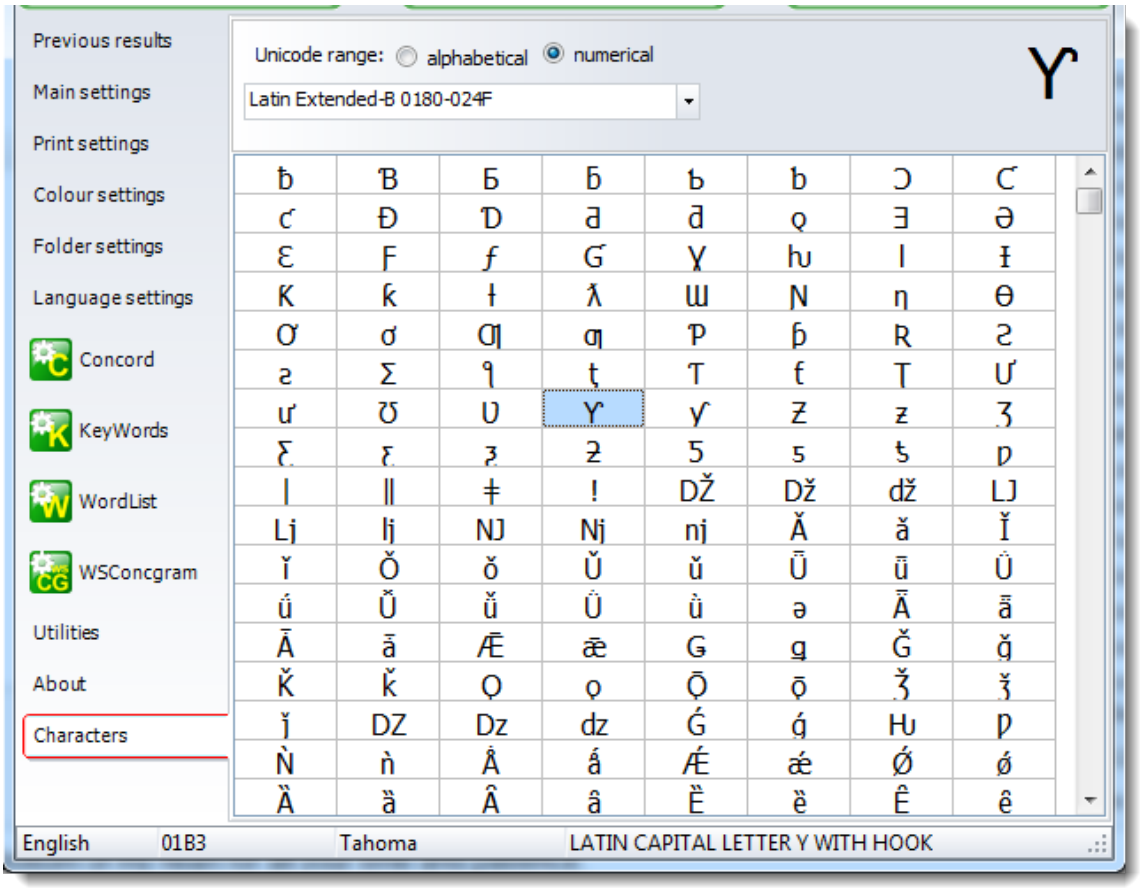
Tip

To check results after changing the code-page, select Choose Texts 52 and choose the menu option *View* for the file in question. If you can't get it to look right, you've probably not got a cleaned-up plain text file but one straight from a word-processor. In that case, take it back into the word-processor (see here 552 for how to do that in MS Word) and save it as text again as a plain text file in Unicode.

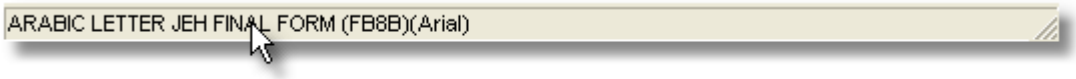
See also: Text Formats 49, Choosing Accents & Symbols 528, Accented characters 527; Choosing Language 77

11.6.2 accents & symbols

When entering your search-word [249] you may need to insert symbols and accented characters into your search-word, exclusion word or context word, etc. If you have the right keyboard set for your version of Windows this may be very easy — if not, just choose the symbol in the main Controller [5] by clicking.



Below, you will see which character has been selected



with the current font (which affects which characters can be seen). You can choose a number of characters and then paste them into Concord, by right-clicking and choosing from the popup-menu:

These options above show Greek, Hebrew, Thai and Bengali characters have been clicked. The last one ("Paste") is the regular Windows paste.

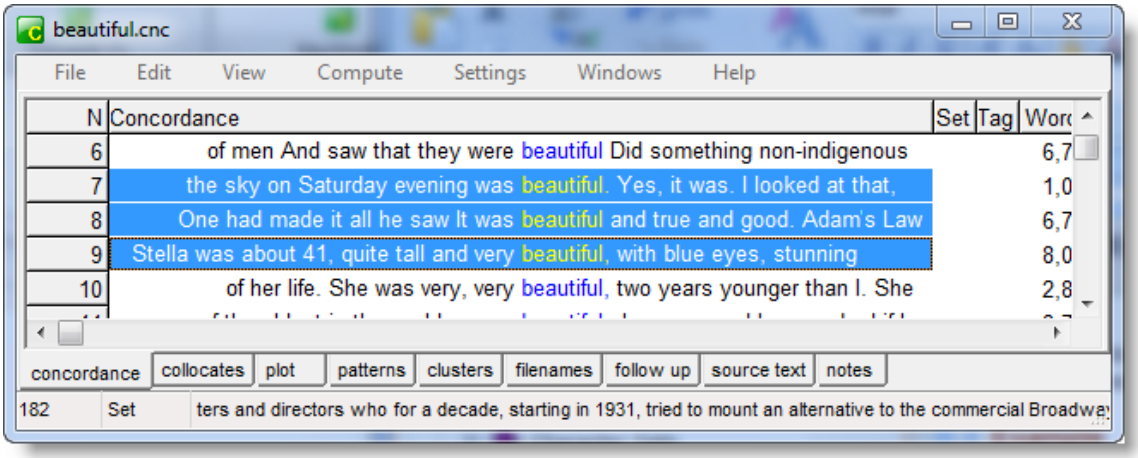
See also: Choosing Language 77, Change Language 526, characters within a word 146

11.7 clipboard

You can block an area of data, by using the cursor arrows and Shift, or the mouse, then press Ctrl+Ins or Ctrl+C to copy it to the clipboard. If you then go to a word processor, you can paste or ("paste special") the blocked area into your text. This is usually easier than saving as a text file 91 (or printing 88 to a file) and can also handle any graphic marks.

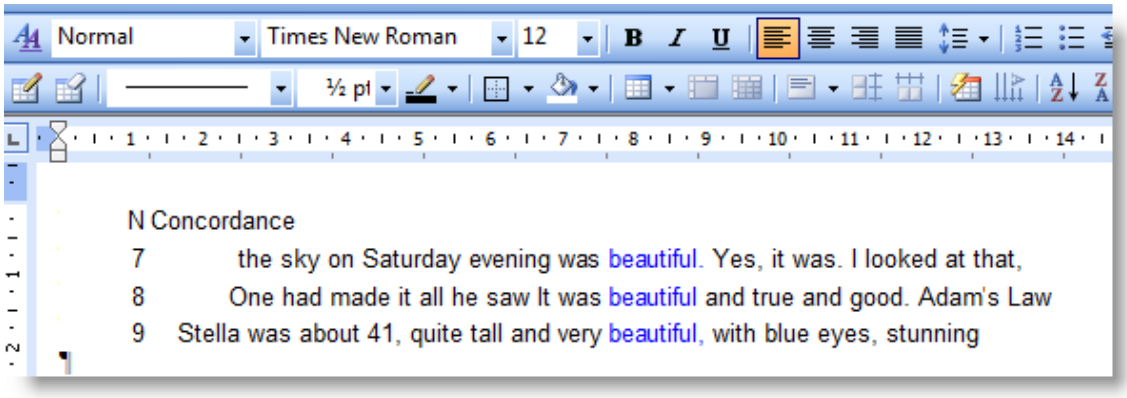
Example

1. Select some data. Here I have selected 3 lines of a concordance, just the visible text, no Set or Filenames information.

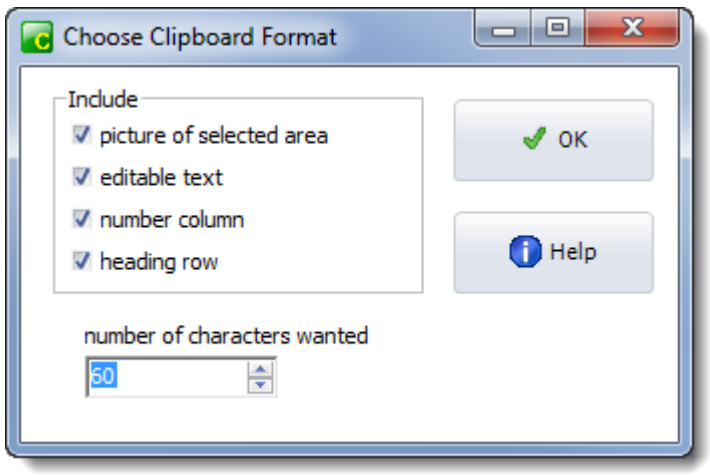


2. Hold down Control and press Ins or C.

In the case of a concordance, since concordance lines are quite complex, you will be asked whether you want a *picture* of the selected screen lines, which looks like this in MS Word:



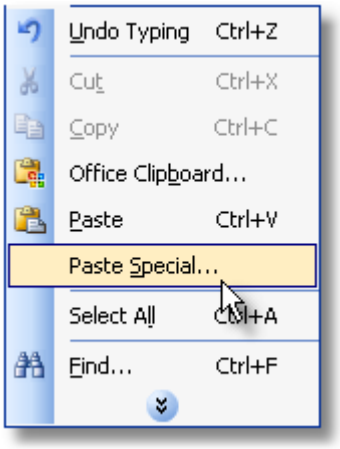
with the colours and font resembling those in WordSmith, and/or plain text, and if so how many characters:



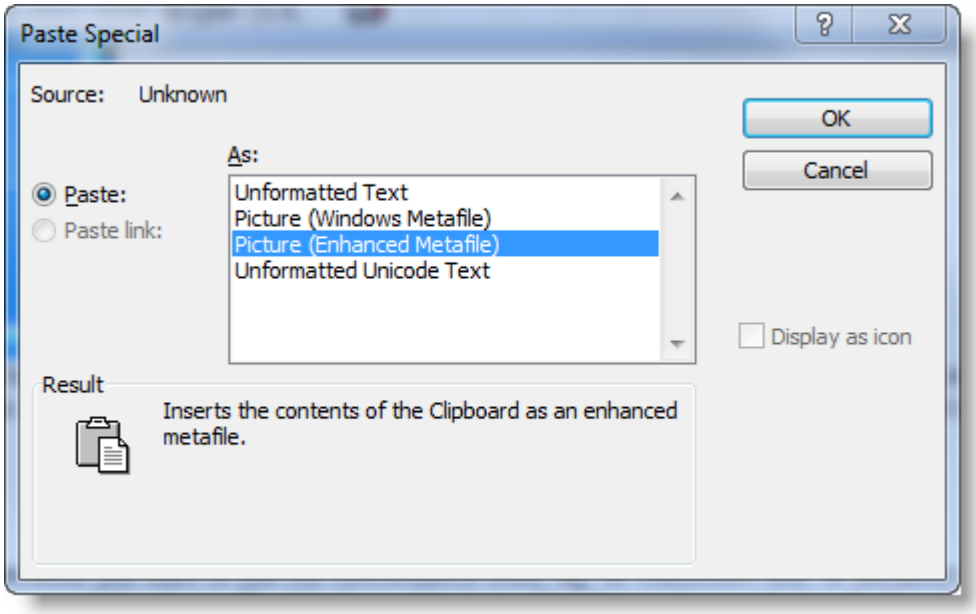
Once you've pressed OK, the data goes to the Windows "clipboard" ready for pasting into any other application, such as Excel, Word, Notepad, etc.

For all other types of lists, such as word-lists, the data are automatically placed in the Clipboard in both formats, as a picture and as text. You can choose either one and they will look quite different from each other!

Choose "Paste Special" in Word or any other application to choose between these formats.



and then, for the picture format

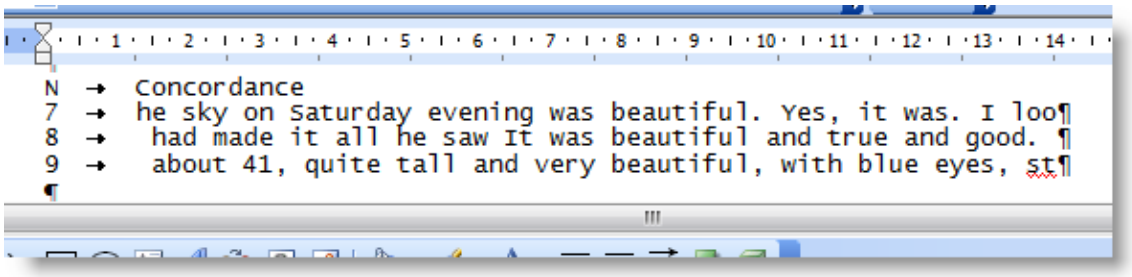


You will probably use this picture format for your dissertation and will have to in the case of plotted data. In this concordance, you get only the words visible in your concordance line (not the whole line).
What you're pasting is a graphic which includes screen colours and graphic data. If you subsequently click on the graphic you will be able to alter the overall size of the graphic and edit each component word or graphic line (but not at all easily!). Note that if you select more lines than will subsequently fit on your page, MS Word may either shrink the image or only paste one pageful.

as plain text

Alternatively, you might want to paste as plain Unformatted Unicode text because you want to edit the concordance lines, eg. for classroom use, or because you want to put it into a spreadsheet such as MS Excel[91]™. Here the concordance or other data are copied as plain text, with a tab between each column. The Windows plain text editor Notepad can only handle this data format. Microsoft Word will paste (using Shift+Ins or Ctrl+V) the data as text. It pastes in as many characters as you have chosen above, the default being 60.

At first, the concordance lines are copied, but they don't line up very nicely. Use a non proportional font, such as Courier or Lucinda Console, and keep the number of characters per line down to some number like 60 or so -- then it'll look like this:



At 10 point text in Lucida Console, the width of the text with 60 characters and the numbers at the left comes to about 14 cm., as you can see To avoid word-wrapping, set the page format in Word to landscape, or keep the number of characters per line down to say 50 or 60 and the font size to 10.

avoid the heading and numbers in WordList or KeyWords too?

See advanced clipboard settings 134.

11.8 contact addresses

Downloads

You can get a more recent version at our website. There are also some free extra downloads (programs, word lists, etc.) there too. And links to sources of free text corpora.

Screenshots

visit https://lexically.net/wordsmith/support/get_started_guides.html for screenshots of what WordSmith Tools can do. This may give you useful ideas for your own research and will give you a better idea of the limitations of WordSmith too!

Purchase

Visit https://lexically.net/wordsmith/purchasing.htm for details of suppliers.

Complaints & Suggestions

Best of all, join Google Groups WordSmith Tools group and post your idea there so others can see the discussion. Or email me (mike (at) lexically.net). Please give me as full a description of the problem you need to tackle as you can, and details of the equipment too. Please don't include any attachments over 200K in size. I do try to help but cannot promise to…

11.9 date format

Date Format

Japanese date format year_month_day_hour_minute. At least it is logical, going from larger to smaller. Why aren't URLs organised in a logical order too?

11.10 Definitions

11.10.1 definitions

valid characters

Valid characters include all the characters in the language you are working with which are defined (by Microsoft) as "letters", plus any user-defined acceptable characters to be included within a word (such as the apostrophe or hyphen 544). That is, in English, **A, a,... Z, z** will be valid characters but **;** or **@** or **_** won't. In Greek, **δ** will count as a valid character. In Thai, (to patak) will be a valid character.

words

The word is defined as a *sequence of valid characters with a word separator 534 at each end*.

A word can be of any length, but for one to be stored in a word list, you may set the length you prefer (maximum of 50 characters) -- any which exceed your limit will get + tagged onto them at that point. You can decide whether or not to include words including numbers (e.g. **$35.50**) in text characteristics 145.

token and type

The term *token* is used to refer to running words and *type* to different words. So in **This is my book, it is interesting** we have 7 tokens but only 6 different types because **is** gets repeated.

clusters

A cluster is *a group of words which are found repeatedly together in each others' company, in sequence*. 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 332 or Concord cluster processing 203 there can be no certainty of this, though clusters often do match phrases or idioms. See also: general cluster information 556.

sentences

The sentence is defined as *the full-stop, question-mark or exclamation-mark (.?!) and (equivalents in languages such as Arabic, Chinese, etc.) immediately followed by one or more word separators 534 and then a number or a currency symbol, or a letter in the current language which isn't lower-case*. Note: languages which do not distinguish between lower-case and upper-case characters do not technically count any as lower case or upper case. (For more discussion see Starts and Ends of Text Segments 164 or Viewer & Aligner technical information 401.)

paragraphs

Paragraphs are user-defined. See Starts and Ends of Text Segments 164 for further details.

headings

Headings are also user-defined -- see Starts and Ends of Text Segments 164.

texts

A text in WordSmith means what most non-linguists would call a text. In a newspaper, for example, there might be 6 or 7 "texts" on

each page. This also means that a text = a file on disk. If it doesn't you're better off totally ignoring the "Texts" column in any WordSmith output.

chargrams

A chargram is a sequence of N consecutive valid characters (excluding digits and punctuation) found in text. e.g. **ABI,ABL,ABO** etc. In English the most frequent 3-chargrams are **THE, ING, AND, ION**.

See also: Setting Text Characteristics 145, Keyness 273, Key key-word 278, Associate 281

11.10.2 word separators

Conventionally one assumes that one word is distinguished from the next by the presence of spaces at either end. But WordSmith Tools also includes within word separators certain standard codes used by most word processors: page eject code (12), tabs (9), carriage return (13) and line feed (10), end-of-text (26). Besides, hyphens 544 may optionally be considered to split words like **self-access** into two words.

Note that in Chinese and Japanese which do not separate words in this way, any WordSmith functions which require word-separation will not work unless you get your texts previously tagged with word-separators.

11.11 demonstration version

The demonstration version of **WordSmith Tools** offers *all* the facilities of the complete suite, except that any screen which shows a list (of words in a word-list, or concordance lines, etc.) is limited to a small number of lines which can be shown or printed. (If you save data, all of it will be saved; it's just that you can't see it all in the demo version.)

See also: Installing 22, Version Information 559, Contact Addresses 532.

11.12 drag and drop

You can get WordSmith to compute some results simply by dragging.

If you have **WordList** open you can simply drag a text file onto it from Windows Explorer and it will create a word-list there and then using default settings. Or if it is not open, drag your text file to the **WordList7.exe** file. Here, *Hamlet* is being dragged onto the WordList tool.

If you have **KeyWords** open you can simply drag a text file onto it from Windows Explorer. If you have a valid word list set as the reference corpus, it will compute the key words.
Or if it is not open, drag your text file to the **`KeyWords7.exe`** file, as in this screenshot where the Dickens novel *Dombey and Son.txt* is being dragged onto the KeyWords file.

If you drag a word-list made by WordList (**`.LST`** ending), a concordance (**`.CNC`**), a key word list (**`.KWS`**) etc. onto the Controller 5, it will open it with the appropriate tool.

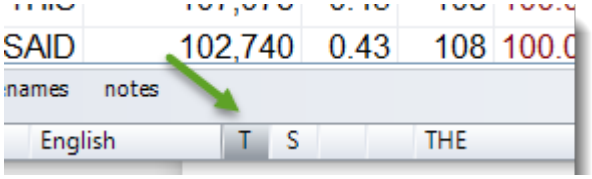
11.13 edit v. type-in mode

Most windows allow you to press keys either
- to edit your data (edit mode), or
- to get quickly to a place in a list (type-in mode)
- to set a category 43.

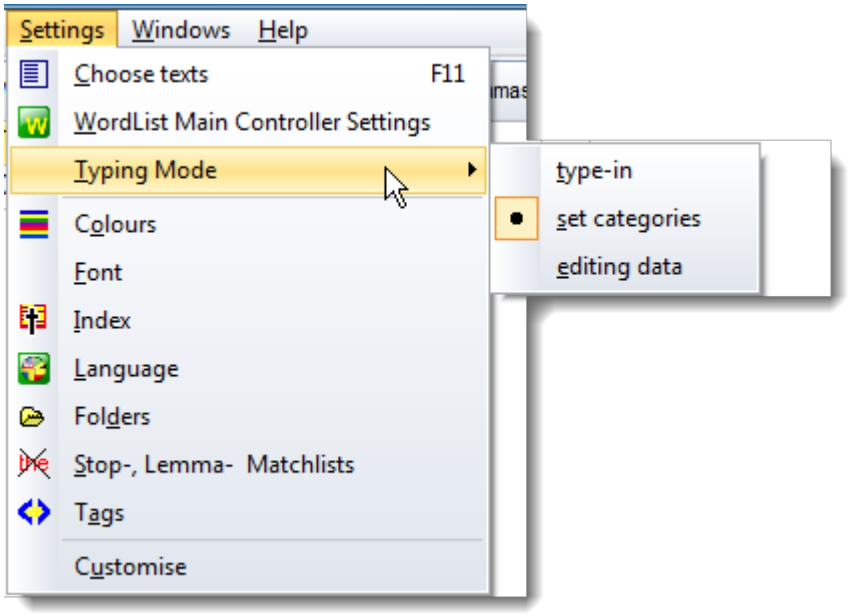
Concordance windows use key presses also blanking 199 out the search word.

In type-in mode, your key-presses are supposed to help you get quickly 106 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 can be done using T (type-in) or S (set) buttons in the status-bar



(here the darker colour **T** button shows the current setting is type-in mode) or via the menu: Settings | Typing Mode:



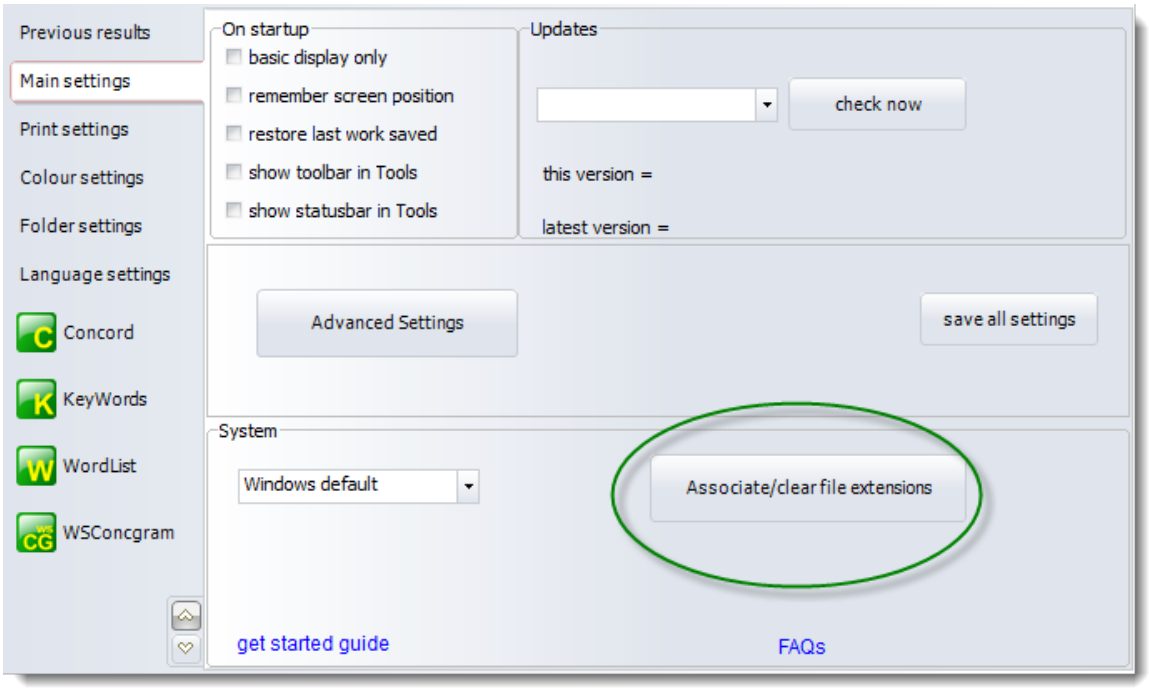
See also: user-defined categories 431.

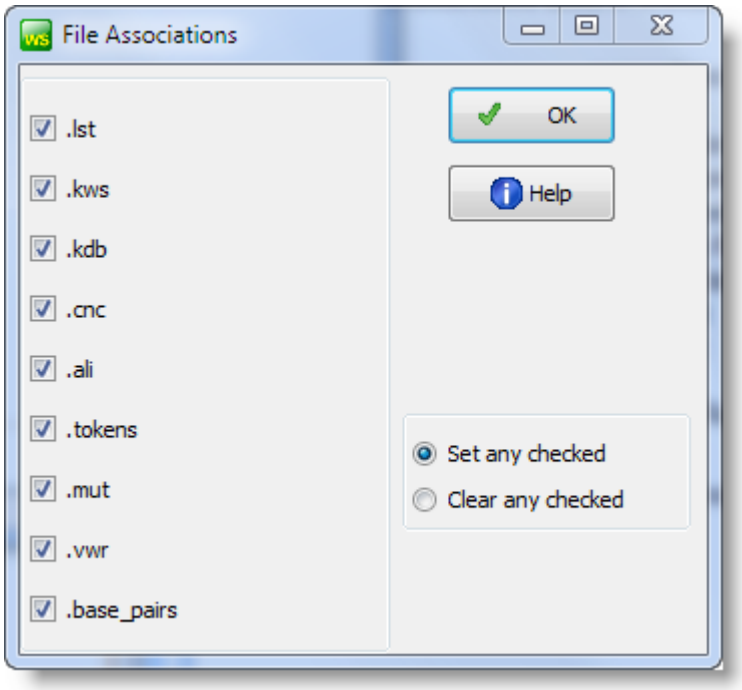
11.14 file extensions

The standard file-extensions used in WordSmith are

.cnc	concordance file
.lst	word list
.mut	mutual information list
.dcl	detailed consistency list
.tokens, .types	word list index file
.kws	key words file
.kdb	key word database file
.base_pairs, .base_index_cg	WSConcgram files
.ali	aligner list
.vwr	viewer list

In the Controller's Main settings, or on installing, you can if you wish associate (or disassociate) the current file-types with WordSmith in the Registry. The advantage of association is that Windows will know what Tool to open your data files with.





11.15 finding source texts

For some calculations the original source texts need to be available. For example, for Concord to show you more context 198 than has been saved for each line, it'll need to re-read the source text.

For KeyWords to calculate a dispersion plot 296, 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 292.

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

If the files are networked ones, WordSmith by default will show them greyed. To detect whether each text file is still available across the network can slow things down. If you do want to have whether files are still present checked, see *Advanced settings | Advanced | Plots, Prompts*.

See also : Source texts 134, Editing filenames 102, Choosing source files 52, find files 320.

11.16 flavours of Unicode

What is Unicode?

What WordSmith requires for many languages (Russian, Japanese, Greek, Vietnamese, Arabic etc.) is Unicode. (Technically UTF16 Unicode, little-endian.) It uses 2 bytes for each character. One byte is not enough space to record complex characters, though it will work OK for the English alphabet and some simple punctuation and number characters.

UTF8, a format which was devised for many languages some years ago when disk space was limited and character encoding was problematic, is in widespread use but is generally *not suitable*. That's because it uses a variable number of bytes to represent the different characters. A to Z will be only 1 byte but for example Japanese characters may well need 2, 3 or even more bytes to represent one character.

There are a number of different "flavours" of Unicode as defined by the Unicode Consortium.

MS Word offers

- Unicode
- Unicode (Big-Endian) (generated by some Mac or Unix software)
- Unicode (UTF-7)
- Unicode (UTF-8)

The last two are 1-byte versions, not really Unicode in my opinion. WordSmith wants the first of these but should automatically convert from any of the others. If you are converting text 489, prefer Unicode (little-endian), UTF16.

Technical Note

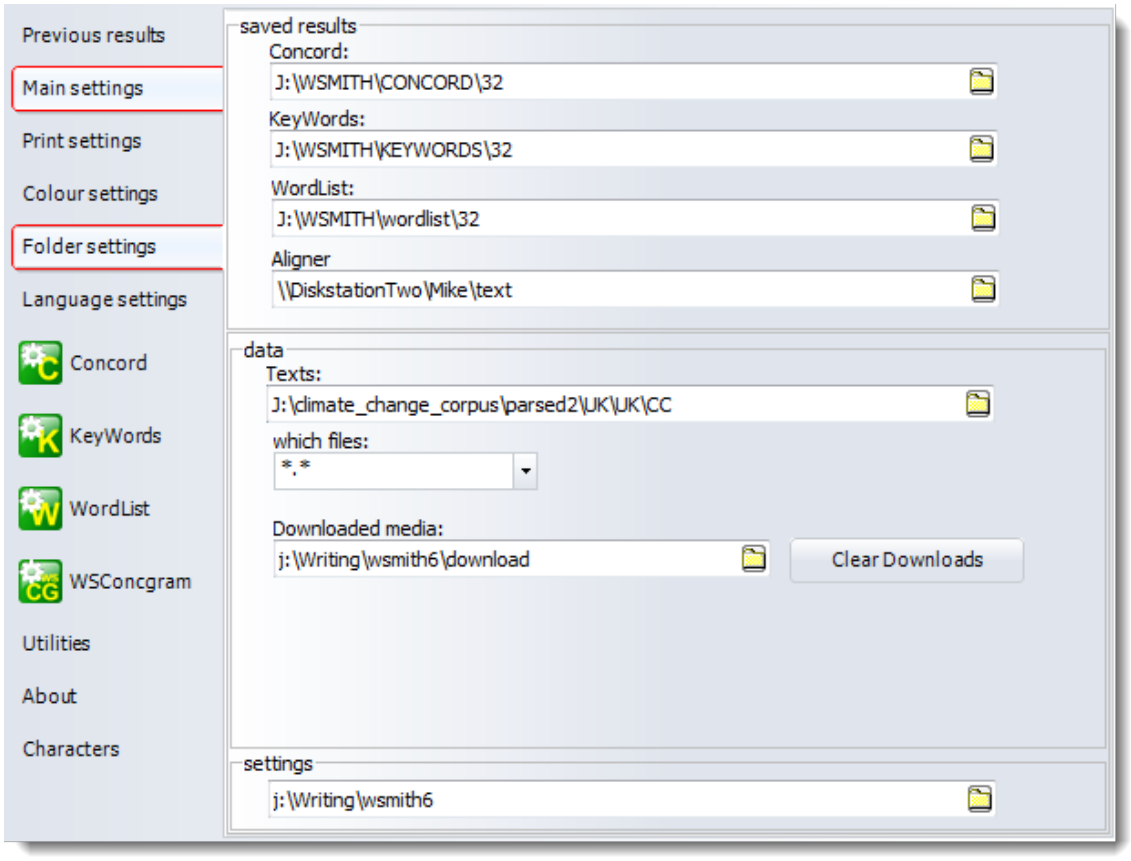
There are other flavours too and there is much more complexity to this topic than can be explained here, but essentially what we are trying to achieve is a system where a character can be stored in the PC in a fixed amount of space and displayed correctly.

Precomposed

In a few cases in certain languages, some of your texts may have been prepared with a character followed by an accent, such as **A** followed by **^** where the intention is for the software to display them merged (**Â**), instead of using precomposed characters where the two are merged in the text file. See the explanation in Advanced Settings 122 if you need to handle that situation.

11.17 folders\directories

Found in main Settings menu in all Tools. Default folders can be altered in WordSmith Tools or set as defaults 110 in **`wordsmith7.ini`**.



- Concordance Folder: for your concordance files.
- KeyWords Folder: for your key-word list files.
- WordList Folder: where you will usually save 91 your word-list files.
- Aligner: for your dual-text aligned work 388
- Texts Folder: where your text files are to be found.

- Downloaded Media: where your sound & video files 258 will be stored after downloading the first time from the Internet.
- Settings: where your settings files (.ini files and some others) are kept.

If you write the name of a folder which doesn't exist, WordSmith Tools will create it for you if possible. (On a network, this will depend on whether you have rights to create folders and save 91 files.)

If you change your Settings folder, you should let WordSmith copy any **.ini** and other settings files which have been created so that it can keep track of your language preferences, etc.

Note: in a network, drive letters such as **G:, H:, K:** change according to which machine you're running from, so that what is **G:\texts\my text.txt** on one terminal may be **H:\texts\my text.txt** on another. Fortunately network drives also have names structured like this: \\computer_name\drive_name\. You will find that these names can be used by **WordSmith**, with the advantage that the same text files can be accessed again later.
If you run WordSmith from an external hard drive or a flash drive 22, where again the drive letter may change, you will find WordSmith arranges that if your folders are on that same drive they will change drive letter automatically once you have saved your defaults 110.

Tip

Use different folders for the different functions in WordSmith Tools. In particular, you may end up making a lot of word lists and key word lists if you're interested in making databases 275 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 537.

11.18 formulae

For computing collocation strength, we can use

- the joint frequency of two words: how often they co-occur, which assumes we have an idea of how far away counts as "neighbours". (If you live in London, does a person in Liverpool count as a neighbour? From the perspective of Tokyo, maybe they do. If not, is a person in Oxford? Heathrow?)
- the frequency word 1 altogether in the corpus
- the frequency of word 2 altogether in the corpus
- the span or horizons 207 we consider for being neighbours
- the total number of running words in our corpus: total tokens

Mutual Information

Log to base 2 of (A divided by (B times C))

where

A = joint frequency divided by total tokens

B = frequency of word 1 divided by total tokens
C = frequency of word 2 divided by total tokens

MI3

Log to base 2 of ((J cubed) times E divided by B)

where

J = joint frequency
F1 = frequency of word 1
F2 = frequency of word 2
E = J + (total tokens-F1) + (total tokens-F2) + (total tokens-F1-F2)
B = (J + (total tokens-F1)) times (J + (total tokens-F2))

T Score

(J - ((F1 times F2) divided by total tokens)) divided by (square root of (J))

where

J = joint frequency
F1 = frequency of word 1
F2 = frequency of word 2

Z Score

(J - E) divided by the square root of (E times (1-P))

where

J = joint frequency
S = collocational span
F1 = frequency of word 1
F2 = frequency of word 2
P = F2 divided by (total tokens - F1)
E = P times F1 times S

Dice Coefficient

(J times 2) divided by (F1 + F2)

where

J = joint frequency
F1 = frequency of word 1 or corpus 1 word count
F2 = frequency of word 2 or corpus 2 word count
Ranges between 0 and 1.

Log Likelihood (different corpora)

where

a = frequency of term 1
b = frequency of term 2

c = total words in corpus 1

d = total words in corpus 2

computes

E1 = c*(a+b) / (c+d) and E2 = d*(a+b) / (c+d)

Log Likelihood is

2*((a* Log (a/E1)) + (b* Log (b/E2)))

(using Log to the base e)

BIC Score

is the log likelihood above - Log(c+d).

Log Likelihood (same corpus)

uses

J = joint frequency
F1 = frequency of word 1 or corpus 1 word count
F2 = frequency of word 2 or corpus 2 word count
T = total word count

then computes K11 = Joint; K12 = F1 * collocation span - Joint; K21 = F2 - Joint; K22 = T - F1 - F2 - Joint

as input to a routine explained at Ted Dunning's blog. The use of the collocation span is proposed by Stefan Evert[524].

Log Ratio

where

a = frequency of term 1

b = frequency of term 2

c = total words in corpus 1

d = total words in corpus 2

computes

Log ((a/c) / (b/d))

(using Log to the base 2)

Conditional Probability (Durrant[524] 2008: 84)

divides the frequency of terms 1 and 2 when together (the joint frequency) by the frequency of term1 (Conditional Probability A) or of term 2 (Conditional Probability B) and multiplies by 100 for better legibility.

Delta Probability (Gries[524], 2013: 144)

where

j = joint frequency of term 1 with term 2

a = frequency of term 1

b = frequency of term 2

c = total words in corpus

computes (j / b) - ((a-j) / (c-a-b+j) and multiplies the result by 100 for legibility. Very similar to Conditional Probability.

Dispersion (Oakes 524 p. 190)

where

n = number of divisions

m = mean of the frequencies over n divisions

sd = standard deviation of the frequencies

v = sd / m

r = square root of n

computes dispersion as 1 - (v / r)

(Oakes suggests square root of n-1 but square root of n gives slightly better results. Either way he says this is designed to range between 1 and 0 but in practice a very low dispersion such as where all the hits are in one division can compute to less than zero. WordSmith will show results of zero or below as blanks.)

Relative Entropy (Gries 524, 2010)

where

n = number of measurements

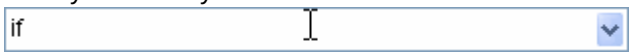
p = the probability of each measurement

computes entropy as the positive sum of (each p * log2 of p) and relative entropy as entropy / log2 of n

See also: link 1 from Lancaster University, link 2 from Lancaster, Mutual Information 342, plot dispersion 554

11.19 history list

History List: many of the combo-boxes in WordSmith like this one for choosing a search-word



remember what you type in so you can look them up by pressing the down arrow at the right.

11.20 HTML, SGML and XML

These are formats for text exchange. The most well known is HTML, Hypertext Markup Language, used for distributing texts via the Internet. SGML is Standard Generalized Markup Language, used by publishers and the BNC; XML is Extensible Markup Language, intermediate between the other two.

All these standards use plain text with additional extra tags, mostly angle-bracketed, such as <h1> and </h1>. The point of inserting these tags is to add extra sorts of information to the text:

1 a header (**<head>**) supplying details of the authorship & edition

2 how it should display (e.g. **<bold>, <italics>**)

3 what the important sections are (**<h1>** marks a heading, **<body>** is the body of the text)

4 how special symbols should display (&eacute corresponds to é)

See also: Overview of Tags 157

11.21 hyphens

The character used to separate words. The item "self-help" can be considered as 2 words or 1 word, depending on Language Settings 145.

11.22 international versions

There are Guides in various languages.

See also: acknowledgements 523

11.23 limitations

The programs in WordSmith Tools can handle virtually unlimited amounts of text. They can read text from CD-ROMs, so giving access to corpora containing many millions of words. In practice, the limits are reached by a) storage 555 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 158 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 171, 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 555 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 544

11.23.1 tool-specific limitations

Concord limitations

You can compute a virtually unlimited number of lines of concordance using Concord.

Concord allows 80 characters for your search-word or phrase 246, though you can specify an unlimited number of concordance search-words in a search-word file 247.

Each concordance can store an unlimited number of collocates with a maximum horizon 207 of 25 words to left and right of your search-word.

WordList limitations

A head entry can hold thousands of lemmas 322, but you can only join up to 20 items in one go using F4. Repeat as needed.

Detailed Consistency 311 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 292-windows per key-word plot 296 display: 20.

number of windows of associates 279 per key key-word display: 20.

File Utilities: Splitter limitations

Each line of a large text file can be up to 10,000 characters in length. That is, there must be an <Enter> from time to time!

Text Converter limitations

There can be up to 500 strings to search-and-replace for each.

Each search-string and each replace-string can be up to 80 characters long.

An asterisk must not be the first or last character of the search-string.

When the asterisk is used to retain information, the limit is 1,000 characters.

If you choose the View option when choosing texts, Notepad will call up the first 10 source text files selected.

See also: General Limitations 544

11.24 links between tools

Linkage with Word Processors, Spreadsheets etc.

All the windows showing lists or texts can easily copy selected information to the clipboard 529. (Use Ctrl+Ins or Ctrl/C 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 https://lexically.net) you can click to access your browser.

Links between the various Tools

The programs in **WordSmith Tools** are linked to each other via wordsmith.exe 5 (the one which says " WordSmith Tools Controller 5 " in its caption, and is found in the top-left corner of your screen). This handles all the defaults 110, such as colours, folders, fonts, stop lists, etc.

In general, if you press Ctrl+C in **WordList** or **KeyWords** you'll go straight to a concordance, computed using the current word and using the current files.

Each Tool will send as much relevant information as possible to the Tool being called. This will include: the current word (the one highlighted in the scrolling window) and the text files where any current information came from.

Example: after computing a word list based on 3 business texts, you discover that the word *hopeful* is more frequent than you had expected. You want to do a concordance on that word, using the same texts. Place the highlight on *hopeful,* hold down Control and press C. Now you can see whether *hopeful* is part of a 3-word cluster 203, or view a dispersion plot.
Example: after computing a key words database 275 using 300 business texts, you discover that the word *bid* seems to be a key key-word, and that it's associated with *company*, *shares* etc. Place the highlight on *bid,* press Control-C and a concordance will be computed using the same 300 texts. Now you can check out the contexts: is *bid* a bid for power, or is it part of a tendering process?
Example: you have a concordance of *green*. Now press Control-W to generate a word list of the same text files. Press Control-K to compare this word list with a reference corpus list to see what the key words are in these text files.

11.25 keyboard shortcuts

scrolling windows:

Control+Home	to top of scrollable list
Control+End	to end of scrollable list
if it's ordered alphabetically:	type-in your search-word 106
and if it scrolls horizontally:	Home -- to left edge End -- to right edge

hotkeys:

Shift-cursor keys	block a section
F1	help
Ctrl+F2	save results 91
F3	print preview
Ctrl+P	print results
F4	join entries
Ctrl+F4	unjoin
Alt+F5	mark entries for joining
Shift+Alt+F5	unmark entry

F5	refresh a list
Shift+Ctrl+F8	auto set row height in Concord
F6	re-sort 372
Ctrl+F6	reverse word sort 372
Shift+Ctrl+F6	word-length sort
F7	view source text
F8	grow line height
Ctrl+F8	shrink
F9	About box (shows version-date and memory availability)
F10	compute collocates
F11	choose texts
Ctrl+Shift+C	compute concordance
Ctrl+C	copy
Ctrl+F3	find again
Alt+D	find next deleted entry
Ctrl+L	layout & columns of data
Ctrl+M	play media file
Ctrl+N	new
Ctrl+U	undo
Ctrl+V	paste
Ctrl+W	close
Alt+X	e**X**it the Tool
Ctrl+Z	Zap 154 deleted lines
Del	delete
Numeric -	delete to the end
Ins	restore deleted entry
Numeric +	restore to the end

see also: Menu items and Buttons 548

11.26 machine requirements

This version of **WordSmith Tools** is designed for machines with:

- at least 1GB of RAM
- at least 200MB of hard disk space
- Windows™ XP or later, or an emulator of one of these if using an Apple Mac or Unix system.

You will find it runs better on a faster 557 machine, especially if there's plenty of RAM 555.
You can run WordSmith from a memory stick 23 on a fast computer better than on a slow computer. (You can run WordSmith on a tiny 10" screen laptop with Windows Starter and little power but all applications on those are slow and there is not much screen for your results.)

There is no Apple Mac version but see https://lexically.net/wordsmith/mac_intel.htm for details on how to use WordSmith on a Mac.

11.27 manual for WordSmith Tools

This help file exists in the form of a manual, which you get when you install 22. The file (**wordsmith.pdf)**, is in Adobe Acrobat™ format. It has a table of contents and a fairly detailed index (which I used **WordList** and **KeyWords** to help me create). Most people find paper easier to deal with than help files!

You may find it useful to see screenshots of **WordSmith** in action: ideas are listed here 532.

11.28 menu and button options

These functions may or may not be visible in each Tool depending on the capacity of the Tool or the current window of data -- the one whose caption bar is highlighted.

advanced

allows access to advanced features

associates

opens a new window showing Associates 279.

auto-join

joins (lemmatises 322) automatically.

auto-size

re-sizes each line of a display so that each one shows as much data as it should. Most windows have lines of a fixed size but some, e.g. in Viewer, allow you to adjust row heights. This adjusts line heights according to the current highlighted column of data.

close (Ctrl+W)

closes a window of data

clusters

computes concordance clusters 203.

collocates

shows collocates 206 using concordance data or recalculates collocates, e.g. after you've deleted concordance lines.

compute

calculates a new column of data 65 based on calculator functions and/or existing data.

column totals

computes totals, min, max, mean, standard deviation for each column 67 of numerical data.

concordance (Shift+Ctrl+C)

within KeyWords, WordList, starts Concord and concordances the highlighted word(s) using the original source text(s).

copy (Ctrl+C)

allows you to copy 69 your data to a variety of different places (the printer, a text file 91, the clipboard 529, etc.).

edit

allows editing 75 of a list or searches for a word (type-in search 106).

exit (Alt+X)

quits a Tool.

edit or type-in mode

alternates between edit and type-in mode.

filenames

opens a new window showing the file names from which the current data derived. If necessary you can edit them 152.

find files

finds any text files which contain all the words you've marked.

grow

increases the height of all rows to a fixed size. See shrink below.

help (F1)

opens WordSmith Help (this file) with context-sensitive help.

join

joins one entry to another e.g. sentences in Viewer, words in WordList (lemmatisation 322).

layout

This allows you to alter many settings for the layout 60: the colour of each column, whether to hide a column of data, typefaces and column widths.

links

computes links[292] between words in a key-words plot.

mark

marks an entry for joining[322] or finding files[104].

match lemmas

checks each item in the list against ones from a text file of lemmatised forms and joins[322] any that match.

match list

matches up the entries in the current list against ones in a "match list file" or template[82], marking any found with (~).

relation

computes mutual information[342] or similar scores in a WordList index list[330].

new... (Ctrl+N)

gets you started[2] in the various Tools, e.g. to make a concordance, a word list, or a key words list.

open... (Ctrl+O)

gives you a chance to choose a set of saved results.

patterns

computes collocation patterns[237].

play media (Ctrl+M)

plays a media file[258].

plot

opens a new window showing a Concord dispersion plot[230] or KeyWords plot[296].

print preview (F3)

previews your window data for printing (Ctrl+P); can print to file, which is equivalent to "save as text[91]".

redo

undoes an undo.

refresh (F5)

re-draws the screen (in Viewer re-reads your text file).

remove duplicates

removes any duplicate concordance[236] lines.

replace

search & replace, e.g. to replace drive or folder data, when editing file-names 152 where the source texts have been moved.

re-sort

re-sorts lists (e.g. in frequency as opposed to alphabetical order) in Concord 238, KeyWords 296 or WordList 372.

ruler

shows/hides vertical divisions in any list; text divisions in a KeyWords plot 296. Click ruler in a menu to turn on or off or change the number of ruler divisions for a plot 230.

save (also Ctrl+F2)

saves your data 91 using existing file-name; if it's a new file asks for file-name first.

save as

saves after asking you for a file-name.

save as text

saves as a .txt file: plain text.

search

searches 95 within a list.

shrink

reduces the height of all rows to a smaller fixed height. See grow above.

statistics

shows detailed statistics 350.

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 141, e.g. proportion of lemmas to word-types.

toolbar

toggles on & off a toolbar with the same buttons on it as the ones you chose when you customised popup menus 120.

undo (Ctrl+U)

undoes last operation.

unjoin

unjoins any entries that have been joined, e.g. lemmatised 322 entries.

view source text

shows the source text and highlights any words currently selected in the list.

Microsoft Excel or Word™

save formatted data for Excel or Word.

wordlist

within KeyWords, makes a word list 293 using the current data.

zap (Ctrl+Z)

zaps 154 any deleted entries.

see also: Keyboard Shortcuts 546, Customising popup menus 120.

11.29 MS Word documents

Inside a **.doc** or **.docx** file there is a lot of extra coding apart from the plain text words. (Actually, a .**docx** doesn't even seem to show the ordinary text words inside it!) For example, the name of your printer, the owner of the software, information about styles etc. For accurate results, WordSmith needs to use clean text where these have been removed.

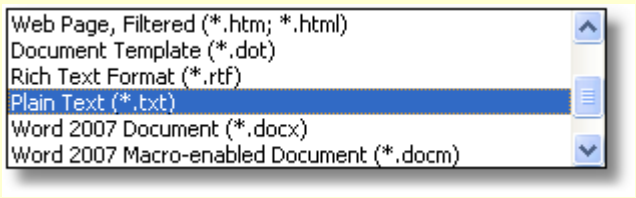
converting your .DOC or .DOCX files

The easiest method, for multiple **.doc** or **.docx** files, is to convert using the Text Converter 489.

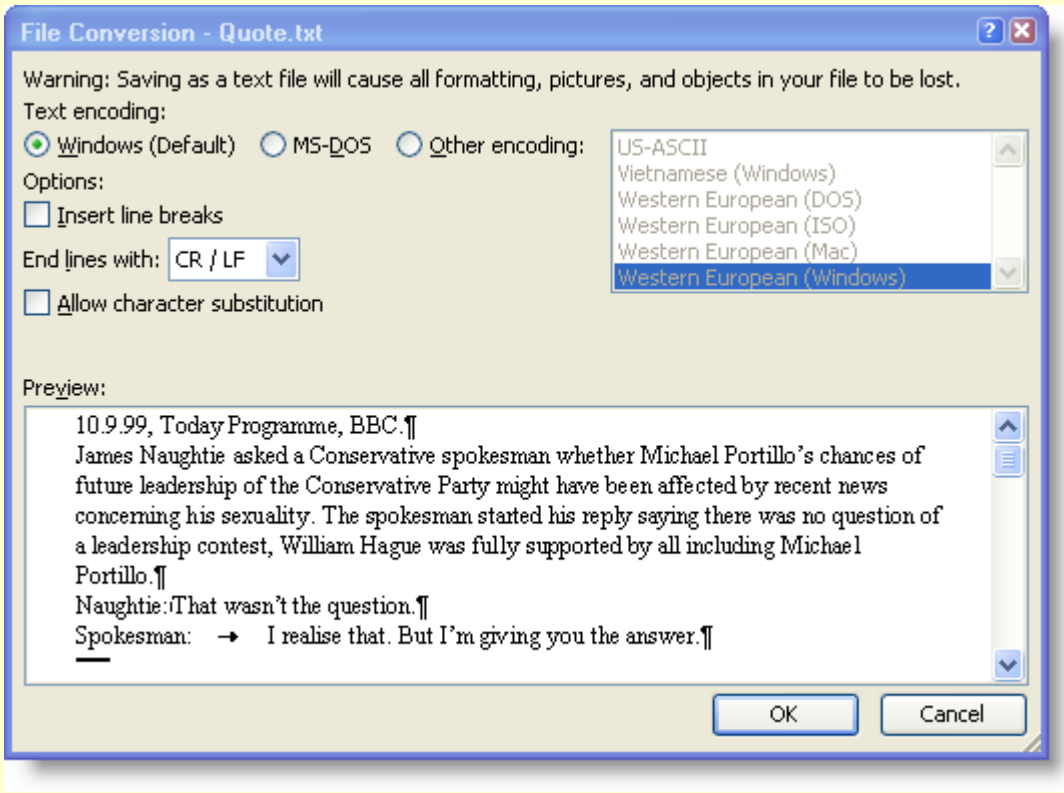
Alternatively you can do it in Word

To convert a **.doc** or **.docx** into plain text in Word can be done thus:

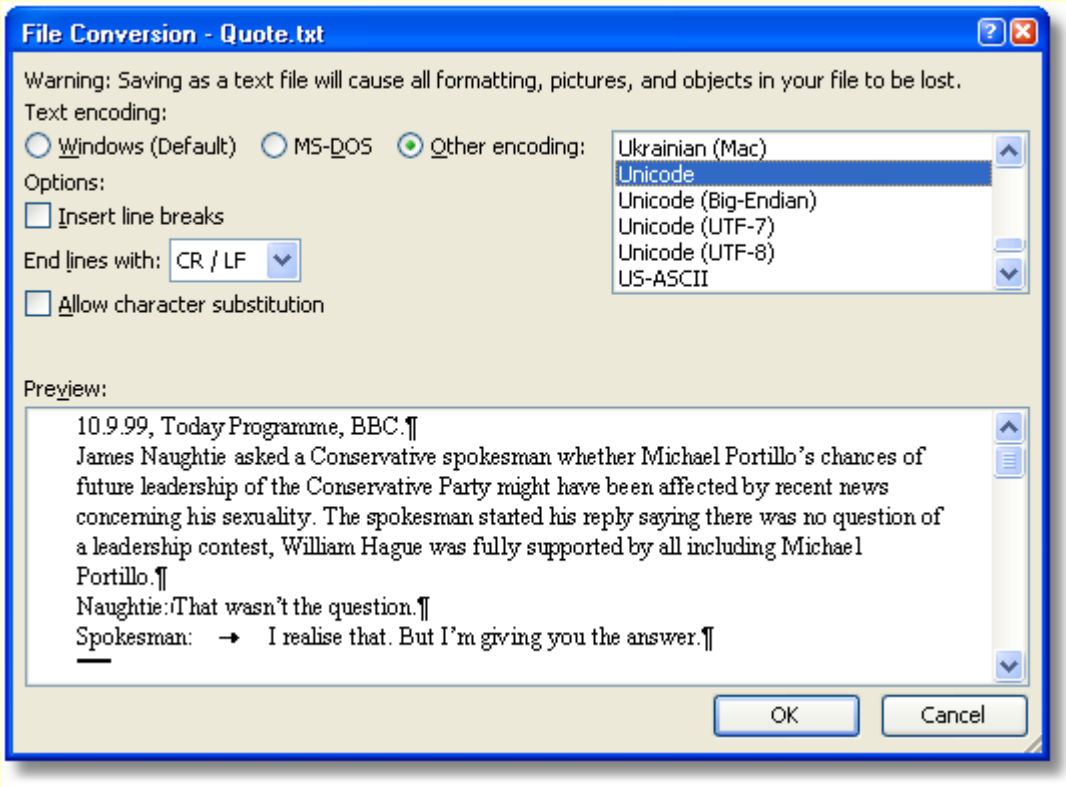
Chose *File* | *Save As* | *Plain text:*



then choose Windows (1-byte per character)



or Other encoding -- Unicode (2-bytes):



11.30 never used WordSmith before

For users who are starting out with WordSmith for the first time, the whole process can seem complex. (After all, the first time you used word-processing software that seemed tricky -- but you already knew what a text is and how to write one...)

So a small text file accompanies the WordSmith installation, and if WordSmith thinks you have never used it before, it will automatically choose that text file for you to start using Concord, WordList etc. WordSmith's method of knowing that you are a new user is

1) have any concordances or wordlists been saved 91?

and

2) has no set of favourite text 59 files been saved for easy retrieval?

11.31 numbers

Depending on Language and Text Settings 145, you might wish to include or exclude numbers from word lists.

11.32 plot dispersion value

The point of it

A dispersion value is the degree to which a set of values are uniformly spread. Think of rainfall in the UK -- generally fairly uniformly spread throughout the year. Compare with countries which have a rainy season.
In linguistic terms, one might wish to know how the occurrences of a word like *skull* are distributed in Hamlet, and WordSmith has shown this in plot form since version 1. The dispersion value statistic gives mathematical support to this and makes comparisons easier.

How it is calculated

The plot dispersion calculated in KeyWords and Concord dispersion plots uses the first of the 3 formulae 540 supplied in Oakes 524 (1998: 190-191), which he reports as having been evaluated as the most reliable.
Like the ruler 548, it divides the plot into (a default of) 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 524, 1996)

See also: KeyWords plot 296, Concord dispersion plot 230, WordList dispersion 354.

11.33 RAM availability

The more RAM (chip memory) you have in your computer, the faster it will run and the more it can store. As it is working, each program needs to store results in memory. A word list of over 80,000 entries, representing over 4 million words of text, will take up roughly 3 Megabytes of memory. (In Finnish it would be much more.) When memory is low, Windows will attempt to find room by putting some results in temporary storage on your hard disk. If this happens, you'll probably hear a lot of clicking as it puts data onto the disk and then reads it off again. You will probably hear *some* clicking anyway as most of the programs in **WordSmith Tools** access your original texts from the hard disk, but a constant barrage of *thrashing* shows you've reached your machine's natural limits.

You can find out how much storage you have available even in the middle of a process, by pressing F9 (the About option in the main *Help* menu of each program). The first line states the RAM availability. The other figures supplied concern Windows system resources: they should not be a problem but if they do go below about 20% you should save results 91, exit Windows and re-enter.

Theoretically, word lists and key word lists can contain up to 2,147,483,647 separate entries. Each of these words can have appeared in your texts up to 2,147,483,647 times. (This strange number 2,147,483,647, half of 2 to the power 32, is the largest signed integer which can be stored in 32 bits and is also called 2 Gigabytes.) You are not likely to reach this theoretical limit: for the item *the* to have occurred 2,147,483,647 times in your texts, you would have processed about 30 thousand million words (1 CD-ROM, containing only plain text, can hold about 100 million words so this number represents some 300 CD-ROMs.) You would have run out of RAM long before this.

If you have a Gigabyte of RAM or more you should be able to have a copy of a word-list based on millions of words of text, and at the same time have a powerful word-processor and a text file in memory.

See also: speed 557

11.34 reference corpus

Reference Corpus

The KeyWords tool compares a study text (or texts) with a comparison set of texts. This can be large and represent some idea of a range of text in a given language, such as the 100 million word British National Corpus. Or it can be chosen as comparable, for example when you study the Election manifesto of one political party and compare it to the manifesto of another or a set of manifestos of several parties.

In the Controller 5 you can set your reference corpus word list 110 for KeyWords and Concord to make use of. (That is, a word list 380 created using the WordList 305 tool.)

11.35 restore last file

By default, the last word list, concordance or key words listing that you saved or retrieved will be automatically restored on entry to **WordSmith Tools**. If the last Tool used is **Concord**, a list of your 10 most recent search-words will be saved too.

This feature can be turned off temporarily via a menu option or permanently in **wordsmith7.ini** (in your Documents\wsmith7 folder).

11.36 single words v. clusters

The point of it…

Clusters are words which are found repeatedly together in each others' company, in sequence. They represent a tighter relationship than collocates, more like multi-word units or groups or phrases. (I call them *clusters* because *groups* and *phrases* already have uses in grammar and because simply being found together in software doesn't guarantee they are true multi-word *units*.) Biber's 524 "lexical bundles" are similar to clusters, require meeting specific thresholds and usually are computed in specific registers or genres (see Conrad & Biber in bibliography 524).

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 206, and semantic prosody (the tendency for **cause** to come with negative effects such as **accident, trouble,** etc.).

WordSmith Tools gives you two opportunities for identifying word clusters, in WordList 332 and Concord 203. They use different methods. Concord only processes concordance lines, while WordList processes whole texts.

How they are computed …

Suppose your text begins like this:

Once upon a time, there was a beautiful princess. She snored. But the prince didn't.

If you've chosen 2-word clusters, the text will be split up as follows:

Once upon
upon a
a time
(note **not** "*time there*" because of the comma)
there was (etc.)
With a three-word cluster setting, it would send
Once upon a
upon a time
there was a
was a beautiful
a beautiful princess
But the prince
the prince didn't
(etc.)

That is, each n-word cluster will be stored, if it reaches n words in length, *up to a punctuation boundary*, marked by **;,.!?** (It seems reasonable to suppose that a cluster does not cross clause boundaries and these punctuation symbols help mark clause boundaries, but there is a Concord setting 249 or a WordList setting 366 for this to give you choice.)

See also: concgrams 511.

11.37 speed

networks

If you're working on a network, WordSmith will be s-l-o-w if it has to read and write results across the network. It's much faster to do your work locally on a **C:** or **D:** drive and then copy any useful results over to network storage later if required.

and generally

To make a word-list on 4.2 million words used to take about 20 minutes on a 1993 vintage 486-33 with 8Mb of RAM 555. 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 144 any further processing if you don't like the results or have enough already). This is a major factor slowing down the processing. Second, reading a file calls on the computer's file management system, which is quite slow in loading it, in comparison with Concord actually searching through it. Third, disk speeds are quite varied, floppy disks being much the worst for speed.
If processing seems excessively slow, close down as many programs as possible and run WordSmith Tools again. Or install more RAM. Get advice about setting Windows to run efficiently (virtual memory, disk caches, etc.) Use a large fast hard drive.
You can run other software while the programs are computing, but they will take up a lot of the processor's time. Shoot-em-up games may run too jerkily, but printing 88 a document at the same time should be fine.

11.38 status bar

The bar at the bottom of a window, which allows you to pull the whole window bigger or smaller, and which also shows a series of panels with information on the current data. The status bar can usually be revealed or hidden using a main menu option. You can right-click on the panel to bring up a popup menu offering choice between Edit, Type and Set 535.

11.39 tools for pattern-spotting

Tools are needed in almost every human endeavour, from making pottery to predicting the weather. Computer tools are useful because they enable certain actions to be performed easily, and this facility means that it becomes possible to do more complex jobs. It becomes possible to gain insights because when you can try an idea out quickly and easily, you can experiment, and from experimentation comes insight. Also, re-casting a set of data in a new form enables the human being to spot patterns.

This is ironic. The computer is an awful device for recognising patterns. It is good at addition, sorting, etc. It has a memory but it does not know or understand anything, and for a computer to recognise printed characters, never mind reading hand-writing, is a major accomplishment.

Nevertheless, the computer is a good device for helping humans to spot patterns and trends. That is why it is important to see computer tools such as these in WordSmith Tools in their true light. A tool helps you to do your job, it doesn't do your job for you.

Tool versus Product

Some software is designed as a product. A game is self-contained, so is an electronic dictionary. A word-processor, spreadsheet or database, on the other hand, is a tool because it goes beyond its own borders: you use it to achieve something which the manufacturers could not possibly anticipate. WordSmith Tools, as the name states, are not products but tools. You can use them to investigate many kinds of pattern in virtually any texts written in a good range of different languages 77.

Insight through Transformation

No, this is not a religious claim! The claim I am making is psychological. It is through changing the shape of data, reducing it and then re-casting it in a different format, that the human capacity for noticing patterns comes to the fore. The computer cannot "notice" at all (if you input 2 into a calculator and then keep asking it to double it, it will not notice what you're up to and begin to do it automatically!). Human beings are good at noticing, and particularly good at noticing visual patterns.

By transforming a text into a list, or by plotting keywords in terms of where they crop up in their source texts, the human user will tend to see a pattern. Indeed we cannot help it. Sometimes we see patterns where none was intended (e.g. in a cloud). There can be no guarantee that the pattern is "really there": it's all in the mind of the beholder.

WordSmith Tools are intended to help this process of pattern-spotting, which leads to insight. The tools in this kit are intended therefore to help you gain your own insights on your own data from your own texts.

Types of Tool

All tools take up positions on two scales: the scale of specialisation and the scale of permanence.

general-purpose ---------- specialised

general-purpose

The spade is a digging tool which makes cutting and lifting soil easier than it otherwise would be. But it can also be used for shovelling sand or clearing snow. A sewing machine can be used to make curtains or handkerchiefs. A word-processor is general-purpose.

specialised

A thimble is dedicated to the purpose of protecting the fingers when sewing and is rarely used for anything else. An overlock device is dedicated to sewing button-holes and hems: it's better at that job than a sewing machine but its applications are specialised. A spell-checker within a word-processor is fairly specialised.

temporary ---------- permanent

temporary

The branch a gorilla uses to pull down fruit is a temporary tool. After use it reverts to being a spare piece of tree. A plank used as a tool for smoothing concrete is similar. It doesn't get labelled as a tool though it is used as one. This kind of makeshift tool is called "quebra-galho", literally branch-breaker, in Brazilian Portuguese.

permanent

A chisel is manufactured, catalogued and sold as a permanent tool. It has a formal label in our vocabulary. Once bought, it takes up storage room and needs to be kept in good condition.

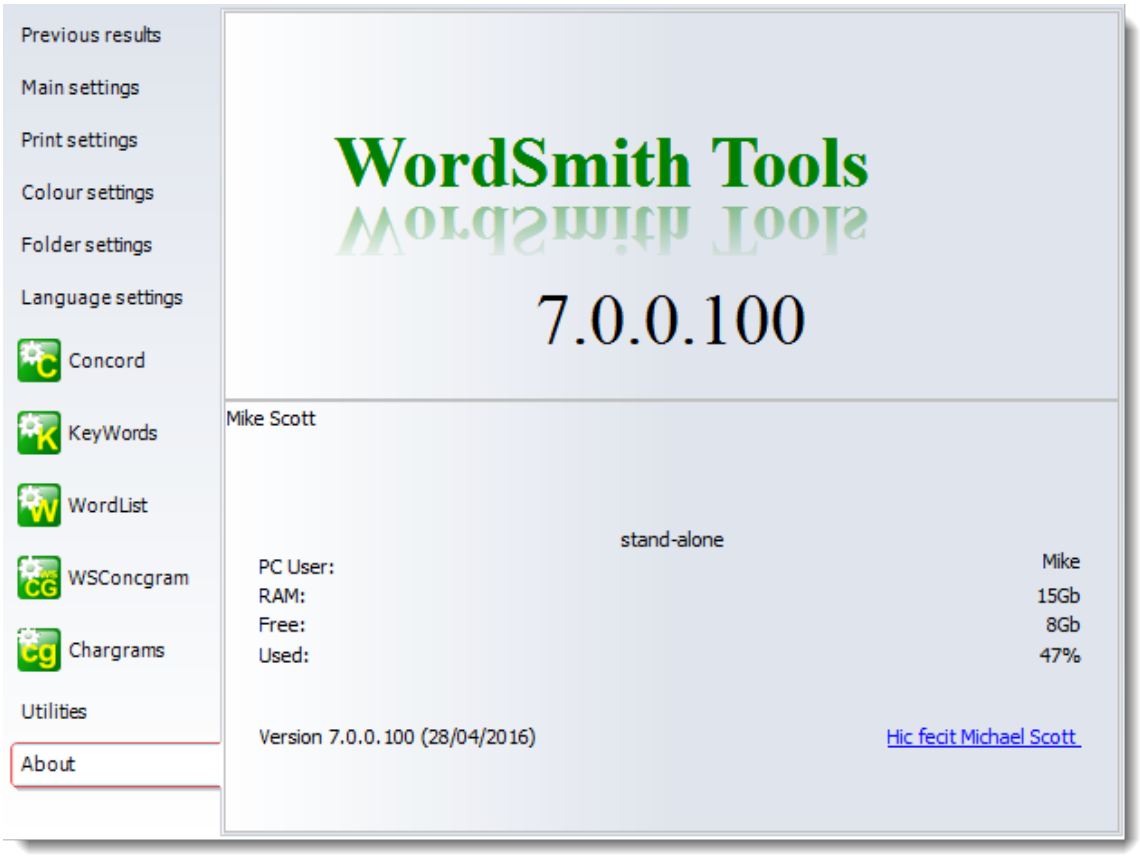
The WordSmith Tools in this kit originated from temporary tools and have become permanent. They are intended to be general-purpose tools: this is the Swiss Army knife for lexis. They won't cut your fingers but you do need to know how to use them.

see also : Word Clouds 152, Dispersion Plots 230, Acknowledgements 523

11.40 version information

This help file is for the current version of **WordSmith Tools.**

The version of **WordSmith Tools** is displayed in the *About* option (F9) which also shows your registered name and the amount of memory 555 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.



See also: 32-bit Version Differences 560, Demonstration Version 534, Contact Addresses 532.

11.40.1 Version 3 improvements

After the earlier 16-bit versions of the 1990s, WordSmith brought in lots of changes "under the hood".

- long file names
- better tag and entity 157 handling including Tag Concordancing 221
- converter for previous data
- zip file handling 561
- easier exporting of data to Microsoft Word and Excel 91
- Unicode text handling, allowing more languages 77 to be processed
- possibility of altering the data 128 as it comes in, e.g. for language-specific lemmatisation
- the old limitations of 16,000 lines of data went. (The theoretical limit for a list of data is over 134 million lines.)

See also: What's New in the current version 4, Contact Addresses 532.

11.41 zip files

Zip files are files which have been compressed in a standard way. WordSmith can now read and write to **.zip** files.

The point of it…

Apart from the obvious advantage of your files being considerably smaller than the originals were, the other advantage is that you don't need to keep thousands of little files you may not have time to examine all of in detail. A batch[30] of thousands of word lists or key word lists can be stored in one **.zip** file.

The "cost" of zipping is the very small amount of time this takes.WordSmith can first unzip it and then show it to you or use the zipped file as input to another process.

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.

65,000 file limit

A **.zip** file cannot hold more than 65,000 different word lists (key word lists, concordances). If you choose more than that a number of **.zip** files will instead get created.

Troubleshooting

12 Troubleshooting

12.1 list of FAQs

See also: logging 120.

These are the Frequently Asked Questions.

There's a much longer list of explanations under Error Messages 569.

Can't process apostrophes 563
Is this Russian, Greek or English? strange symbols in display 564
It crashed 564
It doesn't even start! 566
It takes ages! 566
Keys don't respond 565
Line beyond demo limit 564
Mismatch between Concord and WordList results 564
No tags visible in concordance 563
Printing problem 566
Text is unreadable because of the colours 565
Too much or too little space between columns 563
Wordlist out of order 567
Won't slice pineapples 566

12.2 apostrophes not found

Apostrophes not processed

If your original text files were saved using Microsoft Word™, you may find **Concord** can't find apostrophes or quotation marks in them! This is because Word can be set to produce "smart" symbols. The ordinary apostrophe or inverted comma in this case will be replaced by a curly one, curling left or right depending on its position on the left or right of a word. These smart symbols are not the same as straight apostrophes or double quote symbols.

Solution: select the symbol in the character set in the Controller, then paste when entering your search word 246, or else replace them in your text files using Text Converter 472.

See also: settings 110

12.3 column spacing

column spacing is wrong

You can alter this by clicking on the layout 60 button.

12.4 Concord tags problem

no tags visible in concordance

If you can't see any tags after asking for *Nearest Tag* in **Concord**, it is probably because the Tags to Ignore 158 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 172.

Solution: specify the tag file and run the concordance again.

12.5 Concord/WordList mismatch

Concord/WordList mismatch

If **WordList** finds a certain number of occurrences of a (word list) cluster 556 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 203 it will (by default) take punctuation into account: you can change that in the settings 253 if you wish.

12.6 crashed

it crashed!

Solution: quit **WordSmith Tools** and enter again. If that fails, quit Windows and try again.
Or try logging 132. The idea of Logging is to find out what is causing a crash. It is designed for when WS gets only part of the way through some process. As it proceeds, it keeps adding messages to the log about what it has found & done. When it crashes, it can't add any more messages! So if you examine the log you can see where it was up to. At that point, you may see a text file name that it opened up. Examine that text, you might be able to see something strange about it, eg. it has got corrupted.

12.7 demo limit

demo limit reached

You may have just downloaded, but you haven't yet supplied your registration details. To do this, go to the main WordSmith Tools window, and choose *Settings* | *Register* in the menu.

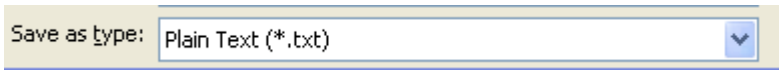
If you haven't got the registration code, contact Lexical Analysis Software (sales@lexically.net). The ***only*** difference between a demonstration version 534 and a full version is: with the latter you can see or print all the data, with the former you'll be able to see only about 25 lines of output.

12.8 funny symbols

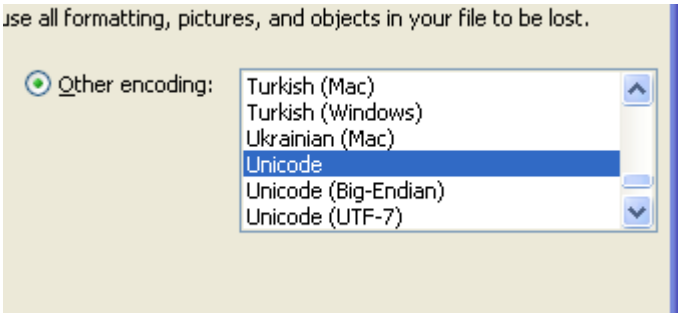
weird symbols

funny symbols when using WordSmith Tools

1. Check your text files. Look at them in **Notepad**. Do they contain lots of strange symbols? These may be hidden codes used by your usual word-processor. Solution: open them in your usual word-processor and *Save As*, with a new name, in plain text format, sometimes called "Text Only" or **.txt**. In Word 2003 the option looks like this:



and then choose Unicode:



2. *Choose Texts*, select the text file(s), right-click and *View*. Does it contain strange symbols?
3. Use Text Converter 489 to clean up and convert and your text files to Unicode.

Greek, Russian, etc.

4. If the text is in Russian, Greek, etc. you will need an appropriate font, obtainable from your Windows cd or via the Microsoft website.
5. If you have several lists open which use *different* character sets, and you change Font 118 or Text Characteristics 145, the lists will all be updated to show the current font and character set, unless you first minimize any window which would be affected.

funny symbols when reading WordSmith data in another application

WordSmith Tools can Save 91 or Save As and Saves as text 91 by printing 88 to a file. "Save" and "Save As" will store the file in a format for re-use by **WordSmith**. This format is not suitable for reading into a word processor. The idea is simply for you to store your work so that you can return to it another day.

"Save as Text", on the other hand, means saving as plain text, by "printing" to a file. This function is useful if you don't want to print to paper from **WordSmith** but instead take the data into a spreadsheet, or word processor such as Microsoft Word. It is usually quicker to copy the selected text into the clipboard 529.

12.9 illegible colours

text unreadable because of colours

Solution: in *Settings*, choose *Colours*. You can now set the colours which suit your computer monitor. Monochrome settings are available.

12.10 keys don't respond

Keys don't respond

If a key press does nothing, it is probably because the wrong window, or the wrong column in the window, has the focus. As you know, Windows is designed to let users open up a number of programs at once on the same screen, so each window will respond to different key-press combinations. You can see which window has the focus because its caption is coloured differently from all the others. The solution is to click within the appropriate window/column, then press the key you wanted.

12.11 pineapple-slicing

won't slice a pineapple

"*Propose to any Englishman any principle, or any instrument, however admirable, and you will observe that the whole effort of the English mind is directed to find a difficulty, a defect, or an impossibility in it. If you speak to him of a machine for peeling a potato, he will pronounce it impossible: if you peel a potato with it before his eyes, he will declare it useless, because it will not slice a pineapple.*" Charles Babbage, 1852.
(Babbage was the father of computing, a 19th Century inventor who designed a mechanical computer, a mass of brass levers and cog-wheels. But in order to make it, he needed much greater accuracy than existing technology provided, and had all sorts of problems, technical and financial. He solved most of the former but not the latter, and died before he was able to see his Difference Engine working. The proof that his design was correct was shown later, when working versions were made. The difficulties he encountered in getting support from his government weren't exclusively English.)

12.12 printer didn't print

printing problem

If your printing comes out with one or more columns printed OK but others blank, you may have pulled your columns too wide for the paper. WordSmith uses information about your printer's defaults to compute what will and will not fit on the current paper. If you can change the printer settings to landscape that will give more space.

12.13 too slow

It takes ages

If you're processing a lot of text and you have an ancient PC with little memory and a hard disk that Noah bought from a man in the market for a rainy day, it might take ages. You'll hear a lot of clicks coming from the hard disk when memory 555 is low. Solution: get a faster computer, by installing more memory which makes a *big* difference), by defragmenting your hard drive, by using a disk cache, or by adjusting virtual memory settings. If you're running **WordSmith Tools** on a network, check with the network administrator whether performance is significantly degraded because of network access.

Solution 2: quit all programs you don't need. That can restore a lot of system memory.

Solution 3: quit Windows and start again. That can restore a lot of system memory.

Solution 4: save and read from the local hard disk (C: or D:), not the network.

12.14 won't start

it doesn't even start

Yikes!

12.15 word list out of order

word-list out of order

Words are sorted according to Microsoft routines which depend on the language. If you process Spanish but leave the Language settings to "English", you will get results which are not in correct Spanish order, (e.g. **LL** will come just before **LM**).
Solution: choose your language 77 and re-compute the word-list.

Error Messages

13 Error Messages

13.1 list of error messages

List of Error Messages

See also: Troubleshooting 563.

Can only save WORDS as ASCII 571
Can't call other Tool 571
Can't make folder as that's an existing filename 571
Can't merge list 571
Can't read file 571
Character set reset to <x> to suit <language> 572
Concordance file is faulty 572
Concordance stop list file not found 572
Conversion file not found 572
Destination folder not found 573
Disk problem: File not saved 573
Dispersions go with concordances 573
Drive not valid 573
Failed to access Internet 573
Failed to create new folder name 573
File access denied 574
File contains none of the tags specified 574
File not found 574
Filenames must differ! 575
Full drive:\folder name needed 575
function not working properly yet 575
INI file not found 570
Invalid Concordance file 575
Invalid file name 576
Invalid Keywords Database file 576
Invalid Keywords file 576
Invalid Wordlist Comparison file 576
Invalid Wordlist file 576
Joining limit reached: join & try again 577
Key words file is faulty 577
Keywords Database file is faulty 577
Limit of 500 file-based search-words reached 577
Links between Tools disrupted 577
Match list details not specified 577
Must be a number 578
Network registration running elsewhere or vice-versa 578
No access to text file: in use elsewhere? 578
No associates found 578
No clumps identified 578
No clusters found 578
No collocates found 579
No concordance entries found 579
No concordance stop list words 579
No deleted lines to Zap 579
No entries in Keywords Database 579
No Key Words found 580

No key words to plot 580
No keyword stop list words 580
No lemma list words 580
No match list words 580
No room for computed variable 580
No statistics available 580
No stop list words 580
No such file(s) found 581
No tag list words 581
Not a valid number 581
No wordlists selected 581
Only X% of reference corpus words found 582
Original text file needed but not found 582
Registration string is not correct 583
Registration string must be 20 letters long 583
Short of Memory! 583
Source Folder file(s) not found 583
Stop list file not found 583
Stop list file not read 583
Tag file not found 583
Tag list file not read 584
This function is not yet ready! 584
This is a demo version 584
To stop getting this annoying message, Update from Demo in setup.exe 584
Too many ignores (50 limit) 584
Too many sentences (8000 limit) 584
Two files needed 584
Truncating at xx words -- tag list file has more! 584
Unable to merge Keywords Databases 584
Why did my search fail? 585
Word list file not found 585
Wordlist comparison file is faulty 585
Word-list file is faulty 585
WordSmith Tools has expired: get another 585
WordSmith Tools already running 585
WordSmith version mis-match 585
xx days left 586

13.2 .ini file not found

.ini file not found

On starting up, **WordSmith** looks for the **`wordsmith7.ini`** file which holds your current defaults 110. If you've removed or renamed it, restore it. This file should be in a sub-folder of your Documents folder called \wsmith7.

13.3 administrator rights

administrator rights

If you see this error message it's because you need Administrator rights to register WordSmith.

Try searching for "Run as Administrator" or this link.

13.4 base list error

base list error

WordSmith is trying to access an word or concordance line above or below the top or bottom of the data computed. This is a bug.

13.5 can only save words as ASCII

Can only save WORDS as Plain Text

WordSmith Tools can't save graphics as a text file. If you get this error message, you can only save this type of data by copying to the clipboard 529 and pasting it into your word-processor.

13.6 can't call other tool

Can't call other Tool

Inter-Tool communication has got disrupted. Save 242 your work, first. Then, if necessary, close down WordSmith Tools altogether, then start the main **`wordsmith7.exe`** program again.

13.7 can't make folder as that's an existing filename

Can't make folder as that's an existing filename

If you already have a *file* called C:\TEMP\FRED, you can't make a *sub-folder* of C:\TEMP called FRED. Choose a new name.

13.8 can't compute key words as languages differ

Can't compute key words as languages differ

Key words can only be computed if both the text file and the reference corpus are in the same primary language. You can compute KWs using 2 different varieties of English or 2 different varieties of Spanish, but not between English and French.

13.9 can't merge list with itself!

Can't merge list with itself

You can only merge 1 word list or key word database with 1 other at a time. Select (by clicking while holding down the Control key) 2 file-names in the list of files.

13.10 can't read file

Can't read file

If this happens when starting up WordSmith Tools, there is probably a component file missing. One example is sayings.txt, which holds sayings that appear in the main Controller 5 window. If you've deleted it, I suggest you use notepad to start a new sayings.txt and put one blank line in it.

If you get this message at another time, something has gone wrong with a disk reading operation. The file you're trying to read in may be corrupted. This happens easily if you often handle very large files. See your Windows manual for help on fragmentation.

13.11 character set reset to <x> to suit <language>

Character set reset to <x> to suit <language>

Prior to version 2.00.07, WordSmith Tools handled fewer character sets 527 and languages 77 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 5, Language Settings | Text, then re-save the list.

13.12 concordance file is faulty

Concordance file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **.CNC, .LST**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **Concord**.

13.13 concordance stop list file not found

Concordance stop list file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.14 confirmation messages: okay to re-read

Okay to re-read?

A confirmation message. To proceed, **Viewer & Aligner** will now re-read the disk file. This will affect any alterations you've already made to the display. You may wish to save first and then try again later.
Also, Viewer & Aligner will try to read the whole text file. If you have a very big file on a slow CD-ROM drive, this will take some time.

13.15 conversion file not found

Conversion file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.16 couldn't save file in folder

WordSmith tried to save a file but either the folder was blocked (read only) or some other file was using the file-name.

13.17 destination folder not found

Destination folder not found

WordSmith couldn't find that folder; perhaps it's mis-spelt.

13.18 disk problem -- file not saved

Disk problem: File not saved

Something has gone wrong with a disk writing operation. Perhaps there's not enough room on the drive. If so, delete some files on that drive.

13.19 dispersions go with concordances

Dispersions go with concordances

They can't be saved 242 separately.

13.20 drive not valid

Drive not valid

WordSmith is unable to access this drive. This could happen if you attempt to access a disk drive which doesn't exist, e.g. drive P: where your drives include A:, C:, D: and E:.

13.21 failed to access Internet

Failed to access Internet

This function relies on a) your having an Internet browser on your computer, b) your system "associating" an Internet URL ending **.htm** with that browser.

13.22 failed to create new folder name

Failed to create new folder or file-name

A folder and a file cannot have the same name. If you already have a *file* called **C:\TEMP\FRED**, you can't make a *sub-folder* of **C:\TEMP** called **FRED**. Choose a new name.
Or you don't have rights to create files in that folder. Or something went wrong while WordSmith was trying to write a file, for example the disk was full up.

13.23 failed to read file

Failed to Read

This may have happened
a) because you included a text file which happens to be empty (zero size), or
b) because your disk filing system has got screwed up, which is especially likely to occur if you often use large files in your word processor (in which do a disk cleanup) or

c) because you tried to use the wrong kind of file for the job (for example the KeyWords procedure won't work if you choose text files as your word-lists).

13.24 failed to save file

Failed to Save

Maybe because you had the same file open in another program or another instance of the Tool you're running. If so, close it and try again.
Or because the folder you're saving to is a read-only folder on a network, or because the disk is full, or because your disk filing system has got screwed up. This last problem is quite common, actually, and is especially likely to occur if you often use large files in your word processor. In that case run *Programs | Accessories | System Tools | Disk Defragmenter*.
If you're working on a network, you will be able to save 242 on certain drives and folders but not others; the solution is to try again on a memory stick or a hard disk drive which you do have the right to save to.

13.25 file access denied

File Access Denied

Maybe the file you want is already in use by another program. You'll find most word-processors label any text files open in them as "in use", and won't let other programs access them even just to read them. Close the text file down in your word processor.

13.26 file contains none of the tags specified

File contains none of the tags specified

You specified tags, but none of them were found.

13.27 file has "holes"

File has "holes"

Text files are supposed to contain only characters, punctuation, numbers, etc. without any unrecognised ones such as character(0). The problem could have arisen because it was transferred from one system to another, part of the disk is corrupted, or else maybe the file contains unrecognised graphics (or else it is not a plain text file but e.g. a Word document) 581.

You can solve this problem by converting the text using the Text Converter. If it is a plain text with holes these will be replaced by spaces. You can find texts with holes using the File Utilities.

13.28 file not found

File not found

This message, like Original Text not found 582, can appear when WordSmith needs to access the original source text used when a list was created, but cannot find it. Have you deleted or moved it? If the file is still available, you may be able to edit the file names in the file name window of this list.

Or the message may come after you've supplied the file name yourself. You may have mis-typed it. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.29 filenames must differ!

Filenames must differ

You can't compare a file with itself.

13.30 folder is read-only

folder is read-only

For some purposes, WordSmith needs to save files e.g. lists of results you have made so that you can get at recent files again. To do this it needs a place where your network or operating system lets you save. Usually \wsmith7 is fine, but in some institutional settings the drive or folder may be "read-only". If you see this message, choose Folder Settings and select there a folder where you can write as well as read.

13.31 for use on X machine only

For use on pc named XXX only

The software was registered for use on another PC. If you get this message, please re-install as appropriate.

13.32 form incomplete

Form incomplete

You tried to close a form where one or more of the blanks needed to be filled in before **WordSmith** could proceed.

13.33 full drive & folder name needed

Full drive:\folder name needed

When typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.34 function not working properly yet

function not working properly yet

This is a function under development, still not fully implemented.

13.35 invalid concordance file

Invalid Concordance file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **.CNC, .LST**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a

.CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **Concord**.

13.36 invalid file name

Invalid file name

File names may not contain spaces or certain symbols such as **?** and *****.

13.37 invalid KeyWords database file

Invalid Keywords Database file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **.KWS, .KDB**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .KDB file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced for a database by the current version of **KeyWords**.

13.38 invalid KeyWords calculation

Invalid Keywords calculation

For KeyWords to calculate the key-words in a text file by comparing it with a reference corpus, both must be in the same language, both must be sorted in the same way (alphabetical order, ascending) and they should both be in the same format (Unicode or single-byte). If you see this message you are trying to compute KWs without meeting these criteria. Solution: open each word-list and check to see it is OK and that it is sorted alphabetically in the same way (in the Alphabetical view, click the top bar to re-sort in ascending alphabetical order), then save it. Check they have both been made with the same language & format settings and if necessary re-compute one or both of them.

13.39 invalid WordList comparison file

Invalid Wordlist Comparison file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **.LST, .CNC**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced as a comparison file by **WordList**.

13.40 invalid WordList file

Invalid Wordlist file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **.LST, .CNC**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .LST file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **WordList**.

13.41 joining limit reached

Joining limit reached: join & try again

Only a certain number of words can be lemmatised 322 in one operation. If you reach the limit and get this message,
1. lemmatise by pressing F4,
2. place the highlight on the head entry again
3. press F5 and carry on lemmatising by pressing F5 on each entry you wish to attach to the head entry
4. when you've done, press F4 to join them up.

13.42 KeyWords database file is faulty

Keywords Database file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .**KDB, .KWS**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .KDB file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced for a database of keywords, by the current version of **KeyWords**.

13.43 KeyWords file is faulty

Key words file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **.KWS, .KDB**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .KWS file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **KeyWords**.

13.44 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 247 to tell Concord to do them in a batch, where the limit is 500.

13.45 links between Tools disrupted

Links between Tools disrupted

WordSmith Tools Controller 5 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 545 are vital in this suite, you should exit WordSmith and re-enter.

13.46 match list details not specified

Match list details not specified

You pressed the Match List 82 button but then failed to choose a valid match list file or else to type in a template for filtering. Try again.

13.47 must be a number

Must be a number

You typed in something other than a number. Be especially careful with lower-case **L** and **1**, and **O** (the letter) instead of **0** (the number).

13.48 mutual information incompatible

Mutual information list is incompatible

A mutual information list derives from an index file, and knows which index file it derives from when computed. Normally when it opens up, it opens up the corresponding index file too. If that index If the file is not found on your PC or has been renamed, you will see this message. The mutual information can still be accessed but a) what you see in terms of Frequency and Alphabetical lists refers to a different index file, and b) it will not be possible to get concordances directly from the listing.

13.49 network registration used elsewhere

Network registration running elsewhere or vice-versa

The site licence registration for use on a network is not valid for use on a stand-alone pc, and vice-versa. If you get this message, please re-install as appropriate.

13.50 no access to text file - in use elsewhere?

No access to text file: in use elsewhere?

The file cannot be accessed. Perhaps another application is using it. If so, close down the file in that other application and try again.

13.51 no associates found

No associates found

Alter settings (*Settings* | *Min & Max Frequencies*) and try again.

13.52 no clumps identified

No clumps identified

Alter settings and try again.

13.53 no clusters found

No clusters found

Alter the settings (*Settings* | *Clusters*) and try again. There were too few concordance lines to find the minimum number needed, or the cluster length was too great.

13.54 no collocates found

No collocates found

In the Controller 5, alter the settings (Concord settings | Min. Frequency) and try again. There were too few concordance lines to find the minimum number needed.

13.55 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 527, check the format of your texts. If you're using a search-word file 247, ensure this was prepared using a plain Windows word-processor such as Notepad.

Have you specified any wildcards 246 (* 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 246 or two. You're more likely to find book* than book.

13.56 no concordance stop list words

No concordance stop list words

13.57 no deleted lines to zap

No deleted lines to Zap

You pressed Ctrl+Z but hadn't any deleted lines to zap 154. No harm done.

13.58 no entries in KeyWords database

No entries in Keywords Database

Alter settings and try again.

13.59 no fonts available

no fonts available for language

The operating system does not have a font which can show the characters for that language. You need to find and install a font.

13.60 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 273 too small for any key words to be detected. For very short texts a minimum frequency of 2 may be needed.

13.61 no key words to plot

No key words to plot

Had you deleted them all?

13.62 no KeyWords stop list words

No keyword stop list words

WordSmith either failed to read your stop-list file or it was empty.

13.63 no lemma list words

No lemma match list words

WordSmith either failed to read your lemma list file or it was empty.

13.64 no match list words

No match list words

WordSmith either failed to read your match list 82 file, or it was empty, or you forgot to check the action to be taken (one option is *None*). Or you tried to match up using a list of words, or a template, when the current column has only numbers. Or else there really aren't any like those you specified!

13.65 no room for computed variable

No room for computed variable

There isn't enough space for the variable you're trying to compute.

13.66 no statistics available

No statistics available

Some types of word list created by **WordSmith Tools**, e.g. a word list of a key words database have words in alphabetical and frequency order but no statistics on the original text files. You cannot therefore call the statistics up in **WordList**. You might also see this message if the statistics file you're trying to call up is corrupted.

13.67 no stop list words

No stop list words

WordSmith either failed to read your stop-list file or it was empty.

13.68 no such file(s) found

No such file(s) found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.69 no tag list words

No tag list words

WordSmith either failed to read your tag file or it was empty.

13.70 no word lists selected

No word lists selected

For **WordSmith** to know which word lists to compare, you need to select them, by clicking on one in each folder. If you've changed your mind, press Cancel.

13.71 not a valid number

Not a valid number

Either you've just typed in, or else **WordSmith Tools** has just attempted to read (e.g. from **`wordsmith7.ini`**, the defaults 110 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 **`wordsmith7.ini`** file for mistakes.

13.72 not a WordSmith file

The file you are trying to open is not a WordSmith Tools file. WordSmith makes files containing your results, files whose names end in **`.LST, .CNC, .KWS`**, etc. These are in WordSmith's own format and cannot be opened up by Microsoft Word -- likewise a plain text file or a **`Word .doc`** 552 cannot usually be read in by WordSmith as a data file, but only as a text file for processing.

13.73 not a current WordSmith file

Not a Current WordSmith File

The file you are trying to open was made using WordSmith but either

- it's a file made using version 1-3
  or
- it's a file made with the beta version of WordSmith and the format has had to change (sorry!)

13.74 nothing activated

Nothing activated

Some forms have choices labelled "Activated" which you can switch on and off. If they are unchecked, you can still see what they would be but **WordSmith** will ignore them.

13.75 Only X% of words found in reference corpus

Only X% of words found in reference corpus

When WordSmith computes key words it checks to see that most of the words in your small word-list are found in the reference corpus, as would be expected. If less than 50% are found, you will get this warning. That is a bit unusual, and is supplied as a warning that for example there might be something strange about one of your two texts. If you know there is nothing strange, then you could ignore the message.

If you are processing clusters you are much more likely to see this warning, however, as the chance of 3-word strings matching in the two lists is less than that of single words matching.

It is up to you to decide whether there is some error in what you are doing or it is OK for many of your smaller word list's words/clusters not to be found in the reference corpus word list.

It might not be so unusual if your reference corpus was very small. But if it is indeed very small, the whole procedure is not very reliable. WordSmith simply looks at the frequencies of each word form and uses basic statistics to compute how greatly they differ in frequency. Basic statistics rely on a notion of what can be expected. If the reference corpus is incredibly small, WordSmith's computation of what is to be expected isn't really very reliable. As a dumb example if you met three citizens of a country you have never visited, and all looked fat, you might suppose the people of that country to be fat in general, but the sample size is not reliable for such an expectation. The KW procedure isn't really proof of anything, incidentally. Words don't occur in texts at all randomly and all ordinary basic statistics can do in my opinion is give us food for thought. So a KW listing isn't proof of anything but it may well give good ideas as to what may prove interesting avenues for research.

13.76 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 152 which the list was based on. But it has been moved or renamed.
Or if on a network, your network connection is not mapped, or the network is down ...or else the right disk or CD-ROM is not in the drive!

13.77 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 88 function in Concord, WordList, KeyWords etc. does need to know what sort of paper size you would print to. If you get a message complaining that no printer has been installed, choose Start | Settings | Printers & Faxes and install a default printer (any printer will do) in Windows.

13.78 registration code in wrong format

Registration code unexpectedly short

PASTE the registration supplied into the box; only paste into the Name or Other Details boxes the details supplied.

If you see this message on registering you may have a registration for a previous major version. If so, contact sales at lexically dot net with your original purchase details and you will be entitled to a 50% discount on the current version.

13.79 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 534 mode.

13.80 short of memory

Short of Memory!

An operation could not be completed because of shortage of RAM 555.

13.81 source folder file(s) not found

Source Folder file(s) not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the filename itself.

13.82 stop list file not found

Stop list file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.83 stop list file not read

Stop list file not read

Something has gone wrong with a disk reading operation. The file you're trying to read in may be corrupted. This happens easily if you often handle very large files, especially if it's a long time since you last ran *Scandisk* to check whether any clusters in your files have got lost. See your DOS or Windows manual for help on fragmentation.

13.84 tag file not found

Tag File not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.85 tag file not read

Tag list file not read

Something has gone wrong with a disk reading operation. The file you're trying to read in may be corrupted. This happens easily if you often handle very large files. See your Windows manual for help on fragmentation.

13.86 this function is not yet ready

This function is not yet ready!

Temporary message, for functions which are still being tested.

13.87 this is a demo version

This is a demo version

You will probably want to upgrade 534 to the full version.

13.88 to stop getting this message ...

Get an update. This is "annoyware" for the demonstration version 534.

13.89 too many requests to ignore matching clumps

The limit is 50. Do any remaining joining manually.

13.90 too many sentences

The limit is 8,000. Do the task in pieces.

13.91 truncating at xx words -- tag list file has more

The tag list file has more entries than the current limit. Or else it isn't a tag list file at all!

13.92 two files needed

You need to select 2 files for this procedure. Select (by clicking while holding down the Control key) 2 file-names in the list of files.

13.93 unable to merge Keywords Databases

Perhaps there wasn't enough RAM 555 to carry out the merge.

13.94 why did my search fail?

The standard search function (F12 or) for a list of data operates on the currently highlighted column. If you want to search within data from another column, click in that column first.

By default, a search is "whole word". Use * at either end of the word or number you're searching for if you want to find it, e.g. in any data consisting of more than one word. (The advantage of the asterisk system is that it allows you to specify either a prefix or a suffix or both, unlike the standard Windows search "whole word" option.)

13.95 word list file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. **.LST, .KWS**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of **WordList**.

13.96 word list file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.97 WordList comparison file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .**LST, .KWS**) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced as a comparison file by **WordList**.

13.98 WordSmith Tools already running

Don't try to start **WordSmith Tools** again if it's already running. Just Alt-tab back to the instance which is running. (You can, however, have several copies of each tool running at once.)

13.99 WordSmith Tools expired

Message for limited period users only. Your version of WordSmith Tools has passed its validity and is now in demo 534 mode. Download another from the Internet 532.

13.100 WordSmith version mis-match

Since the various Tools are linked 545 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 532.

13.101XX days left

Message for limited period users only. At the end of this time WordSmith will revert to demo 534 mode.

Index

- # -

in clusters 332
symbol 380

- . -

.DOC files conversion
  convert to plain text using Text Converter 497
.PDF to plain text 497
.RTF to plain text 497
.XLS to txt 497
.zip files 58

- { -

{CHR(conversion 481

- ~ -

~ operator 226

- 2 -

25 lines 564

- 3 -

32-bit version 560

- 5 -

500 key words 297

- A -

about option 544
accents 528
accents & symbols 528
accessing previous results 87
accurate sort in WordList 372
acknowledgements 523
add a set column 33
add value to corpus 182
adding notes to data 28
Additional requirement 96
adjectives only filter 377
adjust settings 111
adjust timing 261
Administrator rights 570
Adobe .pdf to plain text 497
advanced clipboard settings 134
advanced colour options 116
advanced colour settings 117
advanced concordance settings 249
advanced plot settings 122
advanced scripting 123
advanced settings
  clipboard 134
  convert from UTF-8 120
  force keyboard 120
  Maori 120
  popup menu 120
advanced settings button 113
Aligner
  finding specific words 393
  search 393
alt-tab 151
angle-bracket search span 159
annotate source texts 182
API
  altering your data 128
  custom .dll file 128
  lemmatising with custom .dll 128
apostrophe 146
Apple Mac 548
associate defined 281
associate word-lists and concordances with file-types 536
associated entries
  lemmas 322
associates 279
  follow-up 282
  min. strength 300
  min. texts 300
  statistic 300
asterisks in clusters 336
attach date to text file 56

© 2019 Mike Scott

auto sentence handling 164
auto-joining lemmas 326
automated file-based concordancing 247
automated processing 123

- B -

Babbage 566
batch .zip 30
batch choosing 269
batch concordancing 251
batch processing
  and Excel 30
bibliography 524
BIC score
  formula 540
BIC scores 288
Big5 500
blanking out entries 199
BNC
  handling of sentences and headings 164
  selecting between texts 171
  selecting within texts 168
  tag file 172
boilerplate text 426
boolean and/not 226
breakdowns 138
bugs 525
build a lemma file 444
build sub-corpora 438
burstiness 554

- C -

calculating a plot 296
calling other tools 545
cannot compare word-lists in different languages 576
can't see Concord tags 563
capitalise or lower case conversion 501
CD-ROM
  speed 557
  storage 555
change language of existing data 526
changing font 118
changing from edit to type-in mode 535
Character Profiler
  how to profile text 405
  overview 9
  purpose 404
  settings 410
character sets 527
characters and letters
  accents window 147
  wildcards 148
characters for different languages 528
characters in save as text 253
characters within word 146
chargram
  definition 533
chargram patterns 416
chargram procedure 411
Chargrams
  display 414
  overview 9
  patterns 416
  purpose 411
  settings 418
  syllables 413
Charles Babbage 566
check current version 25
check files exist 122
check lemma consistency 324
check word list 372
checking for updates 114
Cherokee 538
Chinese 538
Chinese Big5 500
Chinese GB2312 500
chm files not visible 25
Choose Languages
  overview 9
choosing reference corpus 274
choosing texts 49
  .DOC files 49
  .DOCX files 49
  .PDF files 49
  .TXT files 49
  clear previous selection 52
  Dickens text 52
  store text files 52
classroom use
  class instructions 59

© 2019 Mike Scott

classroom use
setting up a training sesssion 59
university or school work 59
clipboard 529
clump match 300
cluster
definition 533
cluster settings 366
cluster tag 367
clusters 556
joining 337
reduction & merging 337
clusters in KeyWords 291
Cocoa tags 492
codepages 527
codes 527
codes in search-word 246
collocate follow 213
collocate minimum frequency & length 255
collocates 206
display 210
follow 213
highlighting in concordance 215
horizons 207
lemmas 211
minimum frequency 207
relationships 207
separated by search-word 253
sorting 219
word clouds 219
collocation 217
patterns 237
settings 217
specifications 217
collocation associates 279
collocation breaks 217
colour categories 35
concordances 45
colour categories search 99
colour for ruler 116
coloured tags 174
colouring specific characters 459
colours
changing colours 114
reset colour choices 114
colours of mark-up text 178
column headings 74
column marked green 107
column percentages 131
column tagged conversion 494
column tagged mark-up 494
column totals 67
columns in printing 113
comparing wordlists 308
compute a concordance 69
compute concordance from collocate 213
compute followup scores 300
compute keywords from a word list 307
compute new column of data 65
concgrams 510
filtering 519
generating 513
Concord
blanking 199
breakdowns 200
categories 43
characters to save 242
clusters 203
collocation 206
column visibility 257
creating exercises 199
dispersion 230
export whole concordance line 243
file lists 242
grow 199
hiding tags 257
hits near each other 255
index 195
multiple search-words 247
nearest tag 223
overview 6
path visibility 257
patterns 237
plot 230
printing 242
purpose 195
raw numbers 257
research uses 195
save as 242
saving 242
sentence only 197
sentence only (export) 243
show file data 240
show/hide certain columns 256

© 2019 Mike Scott

Concord
shrink 199
sorting 238
sound and video 258
source text file 199
source text mismatch 242
starting tips 15
stretching the display to see more 199
student use 195
teaching uses 195
text segments 263
text statistics 240
uniform plot 230
view options 196
viewing options 257
what you GET settings 253
what you see and can do 198
what you SEE settings 257
wildcards 246
zapping unwanted lines 236
concordance 196, 198
advice 198
browsing original 198
display 198
grow, shrink 198
highlighting collocates 215
padding 198
purple marks 198
settings 249
concordance batch processing 249
concordance characters 242
concordance characters lining up 529
concordance view options 196
concordances and colour categories 45
concordancing
multimedia 258
tags 221
Concord's save as characters 255
consequence v. consequences 138
consistency analysis (detailed) 311
consistency analysis (simple) 310
consistency lists
sorting 319
contact addresses 532
context horizons 251
context word 226, 251
contextual frequency sort 238
controller (wordsmith exe file) 5
Controller explained 28
convert apostrophes -- curling or straight 491
convert within text files 481
converting BNC XML version 504
converting Treetagger text 494
copy
all 69
choices 69
selective 69
specify 69
copy data to Word 529
copy search-words to Set column 252
corpus
re-arranging 448
restructuring 448
Corpus Checker
aim 419
boilerplate 426
content duplicates 422
corruption detecting 420
name duplicates 425
relevance 429
Corpus Corruption Detector
overview 10
corpus paragraph-count 350
corpus relevance 429
corpus sample 434
corpus sentence-count 350
corpus word-count 350
correcting filenames 102
count data frequencies 141
crash 525
cumulative scores 65
curly quotes 491
custom column headings 74
custom layouts
removing 108
custom processing 128
customising menus 122
cut spaces 253
cutting line starts 168

- D -

data as text file 294

© 2019 Mike Scott

data categories 43
database construction 276
database statistics 275
date format 532
dates of texts 56
deadkeys 122
default colours 114
defaults 110
  .ini files 110
  network 110
defining multimedia tags 162
definition
  tokens 533
  types 533
definition of associate 281
definition of key key-word 278
definition of sentence 533
definitions
  chargrams 534
  clusters 533
  concgram 511
  headings 533
  paragraphs 533
  texts 533
  valid characters 533
  words 533
deleting entries 154
demo limit 564
demonstration version 534
detailed consistency 311
  dice coefficient 319
  relation statistics 319
details of MSWord text 459
dice coefficient
  formula 540
dice coefficient for detailed consistency 319
Dickens text 554
directories 539
dispersion 230
  formula 540
dispersion plot
  sorting 240
dispersion statistic 354
document header 161
DOS to Windows 500
download new version 25
downloaded text problems 459
drag and drop 534
drop a text file onto WordSmith 534
duplicate concordance lines 236
duplicate content 422
duplicate text files
  different names 422
  same names 425

- E -

edit mode 535
editing
  column headings 74
  concordances 236
  delete if 71
  delete to end 71
  random deletion of entries 71
  reduce data to N entries 71
  restore to end 71
  reverse deletion 71
  WordList entries 75
editing your data 70
encrypt your source texts 501
Entitities to characters 492
equal rank words 372
Error messages
  .ini file not found 570
  Administrator rights 570
  base list error 571
  can only save words as ASCII 571
  cannot compute KWs 571
  can't call other tool 571
  can't make folder as that's an existing filename 571
  can't merge list with itself! 571
  can't read file 571
  character set reset to <x> to suit <language> 572
  concordance file is faulty 572
  concordance stop list file not found 572
  confirmation messages - okay to re-read 572
  conversion file not found 572
  couldn't merge KW databases 584
  destination folder not found 573
  disk problem -- file not saved 573
  dispersions go with concordances 573
  drive not valid 573
  expiry date 585

© 2019 Mike Scott

Error messages
failed to access Internet 573
failed to create new folder 573
failed to read file 573
failed to save 574
file access denied 574
file contains "holes" 574
file contains none of the tags specified 574
file not found 574
filenames must differ 575
for use on pc named XXX 575
form incomplete 575
full drive & folder name needed 575
function not working properly yet 575
invalid concordance file 575
invalid file name 576
invalid KeyWords database file 576
invalid KeyWords file 576
invalid WordList comparison file 576
invalid WordList file 576
joining limit reached 577
KeyWords database file is faulty 577
KeyWords file is faulty 577
limit of file-based search-words reached 577
links between Tools disrupted 577
list 569
match list 577
must be a number 578
mutual information 578
network registration used elsewhere 578
no access to text file - in use elsewhere? 578
no associates found 578
no clumps identified 578
no clusters found 578
no collocates found 579
no concordance entries found 579
no concordance stop list words 579
no deleted lines to zap 579
no entries in KeyWords database 579
no fonts available 579
no key words found 580
no key words to plot 580
no KeyWords stop list words 580
no lemma list words 580
no match list words 580
no room for computed variable 580
no statistics available 580
no stop list words 580
no such file(s) found 581
no tag list words 581
no word lists selected 581
not a current WordSmith file 581
not a valid number 581
not a WordSmith file 581
nothing activated 582
only x% of words found in reference corpus 582
original text file needed but not found 582
printer needed but not found 582
read-only folder 575
registration string is not correct 583
registration string must be 20 letters long 583
short of memory 583
source folder file(s) not found 583
stop list file not found 583
stop list file not read 583
tag file not found 583
tag file not read 584
this function is not yet ready 584
this is a demo version 584
too many requests to ignore matching clumps 584
too many sentences 584
truncating at xx words 584
two files needed 584
version mis-match 585
why did search fail? 585
word list file not found 585
word list is faulty 585
WordList comparison file faulty 585
WordSmith already running 585
XX days left 586
example of aligning 389
example of key words 272
examples of tag processing 186
Excel
column totals 91
convert to .txt 497
exclusive search 413
exercises 199
exiting 91
export index data 340
export keywords database data 285
export to spreadsheet etc. 91
exporting concordance
add context word markup 245
add search-word markup 245

© 2019 Mike Scott

external drive folder letters 539
external hard drive 22
extract lemmas from corpus 444
extracting from text files 478

- F -

factory defaults 108, 122
factory settings 110
favourite texts
  loading 59
  saving 59
file associations 536
File Utilities
  compare 2 files 440
  copying files elsewhere 448
  count text files 450
  dates of texts 440
  deleter 450
  dodgy text 459
  editing text file dates 440
  file chunker 458
  find duplicates 422
  find holes in texts 459
  highlighter 442
  index 439
  lemmas 444
  moving files to different folders 448
  overview 10
  rename 451
  Splitter 10
  tags 446
File Utilities Joiner
  joining text files 457
File Utilities Splitter
  bracket first line 454
  end of text separator 454
  end-of-text symbols 457
  filenames 456
  index 453
  purpose 454
  symbols 457
  wildcards 457
File Viewer 459
  overview 10
file-based lemmatisation 326
file-based search-words or phrases 247
filename and path 257
filenames
  display 152
  editing 102
  tab 152
file-types 536
Find All 96
find files containing words 320
find files with KWs 104
find which files contain a word or cluster 320
Find window 96
finding a word
  word or part of a word 95
finding by typing 106
finding relevant files 104
finding source texts 537
first use of WordSmith 554
fixed columns 64
flash drive folders 539
folder letters 539
folder settings 118
folder view 113
folders 539
  force detailed view 113
folders created using text converter 477
follow-up concordancing 45
fonts
  greek 118
  russian 118
force folders to show in detailed view 113
format test for text files 54
formulae
  BIC score 540
  Conditional probability 540
  Delta probability 540
  Dice coefficient 540
  dispersion 540
  Log Likelihood 540
  Log Ratio 540
  MI 540
  MI3 540
  T-score 540
  Z-score 540
freeze columns 64
frequencies of suffixes 359
full lemma processing 208, 299

© 2019 Mike Scott

- G -

GB2312 500
general settings 113
get favourite text selection 59
getting started 2
getting started with Concord 15
getting started with KeyWords 16
getting started with WordList 18
globality of plot 554
green marking in left column 107
grow a concordance line 199
grow and shrink 198
grow/shrink concordance lines 199

- H -

handling hypens 146
handling multiple windows 151
handling of numbers 146
handling Word .doc files 552
hash representing words with numbers 380
header 161
header removing 167
headings
  definition 533
  start & end 164
hex 459
hide number columns in Concord 256
hide tags 224, 253
hide words 253
highlight words in a text 442
highlighting 442
highlighting collocates in concordance 215
history list 87
holes in file 574
horizons 207
hotkey combinations 546
hotkeys
  Ctrl/F2 91
  list 546
  space bar 199
how many words 544
how much text 544
how to build a database 276
HTML 543
HTML headers
  cutting out 167
HTML/BNC/XML entities to characters 492
hyphen treatment 544
hyphens 146

- I -

idioms 556
ignore punctuation 255
illegible 565
importing text into a word list 363
incompatibility between word lists 286
index lists
  export 340
  making Wordlist Index 330
  uses 330
  viewing 338
index relationships 344
index settings 366
information about WordSmith version 559
installing WordSmith Tools 22
instructions folder 25
interface 544
international versions 544
introduction to WordSmith Tools 2
inverted commas 563
it won't do what I want 563

- J -

Japanese 538
Japanese ShiftJis 500
joining clusters 337
joining entries 322

- K -

key key word defined 278
key key-words 275
key word database settings 300
key word procedure setting 297
key word settings in Controller 297
keyboard 546
keyness

© 2019 Mike Scott

keyness
definition 273
p value 273
thinking about 274
key-ness defined 273
keyness scores 275
keyword database
min. frequency 300
min. Kws per text 300
related clusters 285
settings 300
statistics 275
KeyWords
advice 286
associates 281
calculation 287
choosing your files 269
clusters 291
column visibility 300
compute a word list 293
database 275, 276
database exporting 285
display 302
example 272
failure/problems 286
finding KWs in other texts 104
index 266
key key-words 278
lemma processing 299
links 292
overview 7
plot 296
purpose 266
sorting 296
starting tips 16
statistics 287
thresholds 287
tips 286
what you GET settings 297
what you SEE settings 300
keywords minimal processing 297
KeyWords plot
display 294
Korean Hangul 500
KWs in other text files 104

- L -

language
Baltic 77
Central European 77
change language in saved data 526
Cyrillic 77
Greek 77
Portuguese 77
Russian 77
language settings
characters within a word 145
end of heading marker 145
end of paragraph marker 145
end of sentence marker 145
heading marker 145
headings (specifying) 145
hyphens 145
numbers 145
Languages Chooser
apostrophe 79
font 81
language 79
new language 82
saving settings 81
Languages Chooser: overview 80
layout
alignment 60
column width 60
decimal places 60
editing column headings 60
grid format 60
headings 60
save layout 60
typeface 60
lemma absent from a list 329
lemma consistency 324
lemma file 327
building 444
lemma list 327
lemma visibility settings 369
lemmas 321
auto-joining 326
file 327
file-based 326
in KeyWords 299

© 2019 Mike Scott

lemmas 321
joining automatically 326
joining manually 325
matching in WordList 327
template 326
texts column has 0 324
visibility 369
lemmatising source texts 501
lemmatising using a template 326
letter-count 350
lexical bundles 556
licence details 24
limitations 544
links between tools 545
list of menu options 548
list text files meeting criteria 122
listing all the plot marks 294
localisation 544
log file to trace problems 132
log likelihood 288
formula 540
log likelihood computing 344
Log Likelihood score 342
Log ratio procedure 288
logging 132
lowest possible value for clusters 332
LY endings in a word list 359

- M -

Mac version 548
machine requirements 548
Main requirement 96
make a word list from keywords data 293
manual for WordSmith Tools 548
manual joining 325
manual lemmatisation 325
mark_up 159
custom settings 165
document header removal 159
entity references 159
tags as selectors 159
types of 159
marking 325
marking context word in txt export 245
marking entries
green margin 108
unmarking 108
white margin 108
marking MWUs 482
marking search word in txt export 245
mark-up 157
autoload tag file 158
colours 178
handling tag-types 158
HTML & SGML tags 158
making a tag file 172
multimedia 162
section tag 172
selecting between texts 168, 171
selecting within texts 168
video 172
match list
filtering 82
mark words in a word list 82
max followup items 300
mean and standard deviation 350
media timing 261
memory stick 22, 23
memory usage 555
menu choices 548
menu shortcuts 122
merge concordances 310
merge wordlists 310
MI score 342
MI3
formula 540
MI3 computing 344
MI3 score 342
Microsoft Word 529, 552
Minimal Pairs
aim 462
choosing files 464
output 466
overview 11
requirements 462
rules and settings 467
running the program 468
minimal processing (Concord) 255
minimal processing (KeyWords) 299
mismatch in file lists 242
mismatched tags 446
modify source texts 182

© 2019 Mike Scott

moving/copying files elsewhere 448
MS Word 552
multimedia tags 162
multiple file analysis 275
multiple lists 30
multi-word unit 182
multi-word-unit marking 482
mutual information
  computing 344
  formula 540
mutual information scores 342
mutual information screen 342
mwu tag 488

- N -

n/a 384
nag message 584
nearest tag 223
negative keyness 273
negative keywords 297
negative KWs
  example 302
network folders 122
network settings 25
network speed 557
network version 25
networks
  defaults 110
new in this version 4
new user 554
n-grams 556
n-grams in WordList 332
no hits within X words 255
non-characters allowed within a word 146
normalisation 131
normalised counts 131
notes 28
number handling 146
number of concordance entries 253
number of texts and lemmas 324
number sort 238
number then paragraphs in your corpus 481
numbering paragraphs
  Viewer & Aligner 395
numbering sentences
  Viewer & Aligner 395
numbers
  how treated 554
numbers in word lists 146
numbers in words: display 380

- O -

obtaining video and sound files 263
omit # in clusters 332
online screenshots 4
only if containing 171
only words with tag-file tags 187
options for defaults 110
over-writing 473

- P -

p value 273
padding out the search-word with space 257
paragraph
  start & end 164
paragraph marker 145
paragraph numbers
  Text Converter 481
paragraphs
  definition 533
  specifying 145
parallel concordance 393
part of speech (POS)
  tags 374
parts of speech 172
pasting
  as graphic or as text 529
  concordance into Word 529
  paste special 529
patterns
  highlighting in concordance 215
pen drive 22, 23
per 1000 words 131
per million words 131
percent column 131
percentages v. raw numbers 150
permanent settings 110
phrase frames 336
phrases 556

© 2019 Mike Scott

phrases in WordList 306
plain text 49
plot context settings 86
plot contexts 86
plot details listed 294
plot dispersion calculation 554
plot dispersion value 554
plot divisions 116, 117
plot marks
  list 294
plot with grouped files 234
plots and links 292
plotting key words 296
POS mark-up 172
POS tags 374
precomposed characters 122
prefix frequencies 359
prefix tags 374
previous lists 87
print
  blank page 89
  preview 89
print preview
  zoom 89
printer settings 113
printing
  footer 88
  header 88
  landscape 88
  portrait 88
process text file if it contains X 171
prompts to save 122
punctuation breaks 217
purple marks in word list display 380

- Q -

quitting 91
quotation marks 563

- R -

RAM availability 555
randomised concordance entries 253
range 310, 311
raw numbers 150, 257
raw numbers v. percentages 150
recompute file lists 242
re-compute filenames after zapping 154
recompute token count 349, 366
recomputing plot 234
reference
  .DOC files 552
  API 523
  Application programming interface 523
  convert to .TXT in MS Word 552
  programming WordSmith 523
reference corpus 555
registering WordSmith 471
registration 471
registry 536
relationship between collocate and search-word 207
relationship computing
  Log Likelihood 342
  MI statistic 342
  MI3 statistic 342
  T score 342
  Z score 342
relationships computed from an index 344
relationships screen 342
relationships: case sensitivity 208
relevance check
  filter relevant texts 434
  RTF sample 436
relevance filter 429
removable drive 23
remove all mark-up from a corpus 492
remove custom layouts 108
remove duplicates 236
remove line-breaks 489
remove messages 108
remove some XML mark-up 495
rename numerous files 451
re-ordering 154
re-ordering word lists 75
repeated concordance lines 236
repetition
  boilerplate 426
replacing 473
report on a crash 525
re-sorting
  collocates 219

© 2019 Mike Scott

restore all defaults 110
restore factory defaults 122
restore factory settings 110
restore last file 555
restore last work 113
restore settings 108
restricted search 226
ruler 294
running words 533

- S -

save as
  Excel 91
  HTML 91
  text 91
  XML 91
save favourite text file set 59
save prompts 122
save text file data 122
saving
  Ctrl/F2 91
  partial 91
saving defaults 110
scripts 123
search & replace 102
search by typing 535
search span 159
search text 101
searching 95
searching by typing 106
search-word
  advanced functions 249
  alternative search words 246
  ascii codes for searching 246
  asterisk 246
  boolean or 246
  case sensitivity 246
  CHR 246
  file-based 247
  history list 246
  slash 246
  syntax 246
  text file specifying 247
  whole word 246
section
  start & end 164
selecting between texts 168
selecting multiple entries 107
selecting within texts 168
sentence
  auto handling 164
  start & end 164
sentence breaks 217
sentence lengths exporting 340
sentence marker 145
sentence only 196, 253
  exporting (zero characters setting) 243
sentence only view in Concord 197
sentences
  definition 533
  specifying 145
separate search-words 253
Set column 43
  adding 33
  removing 33
  which is active 33
set column colours 35
set textual date 56
settings 110
  advanced 113
  colours 114
  defaults 110
  folders 118
  fonts 118
  general 113
  language 77, 145
  main 113
  permanent 110
  printer 113
  restoring 108
  show help 113
  statusbar 113
  toolbar 113
ShiftJis 500
shortcuts 546
show BIC in plots 297
show descriptions 178
show file data 240
show help at startup 110
show help file 113
show or hide data below a minimum threshold 60
show or hide tags 224
show text file

© 2019 Mike Scott

show text file
highlight KWS 432
shrink a concordance line 199
SI numbers 146
single words 556
slow 566
sorting
Concord 238
consistency lists 319
dispersion plot 240
KeyWords 296
tags 223
word list 372
sound & video tagged files 258
sound file tags 162
source text(s)
view 134
source texts 537
converting to a better format 489
modify 182
speed 557
SRT files
conversion 502
obtaining 263
transcripts 263
standardised or mean type/token ratio 353
start and end of sentence 164
statistics
headings 350
letters 350
paragraphs 350
sections 350
sentences 350
word list statistics 350
word-count 350
status bar 548, 558
statusbar 113
stop at punctuation 217
stop at sentence break 217
stop lists 141
stop lists v. match lists 362
stoplist.cod 484
stopping 144
storage 555
strange characters 49
STTR 353
suffix frequencies 359
summary statistics 138
summary statistics (general) 141
suspending processing 144
swap tags and words 492
syllable-chargrams 413
symbols 528

- T -

T score 342
tag colours 178
tag concordancing 221
tag file
colours 178
descriptive label 172
video 172
tag filtering 377
tag renaming 178
tag span 158
tag string only tags 179
tag structure 172
tag types 159
tag visibility 224
tag-free corpus 492
tagged text 157
tags
finding mis-matched 446
overview 157
span 158
tags and attributes 172
tags but no words 188
tags for video 172
tags in WordList 374
tags swapped with words 492
tags to exclude 172
tags to retain 172
teacher instructions 59
TED talks 263, 502
templates 426
Test for Unicode 54
test text file format 54
text characteristics 145
Text Converter 489
.DOC 497
.DOCX 497
.Excel 497

© 2019 Mike Scott

Text Converter 489
.PDF 497
.RTF 497
<%IFDEF NEXT_TOPIC%>multi-word linking<%ENDIF%> 11, 472, 473, 478, 479, 484, 485, 486, 489, 497, 504
asterisk 485
BNC XML version conversion 504
change case 504
clusters 488
conversion file 484
cutting header 473
extracting 478
folders 473
index 472
insert numbering 481
into Unicode 489
just one change 481
line-breaks removal 489
move if 479
multi-word units 481
overview 11
phrases 488
removing all tags 485
sample conversion file 486
settings 473
syntax 485
Unicode conversion 489
UTF16 conversion 489
UTF8 conversion 489
wildcards 485
text date analysis 149
text dispersion key words 287
text file
use to build a word list 363
text file dates 440
text format for WordSmith 49
text formats 145
text segments in Concord 263
texts
choosing 52
favourites 59
more texts 52
texts associated with a kw 282
tie-breaking 238
time-lines 149
title text 179
TMX files 499
to right only 346
token
definition 533
token count 349
token recomputing 366
toolbar 113, 548
tools for pattern-spotting 558
tool-specific limitations
Concord 544
KeyWords 544
Text Converter 544
Viewer & Aligner 544
WordList 544
training students 59
translation memory files 499
treatment of hyphens 146
Treetagger conversion 494
triangle 384
Troubleshooting 563, 566
accented symbols 564
apostrophes not found 563
colours unreadable 565
column spacing 563
Concord tags problem 563
Concord/WordList mismatch 564
crashed 564
curly quotation marks 563
demo limit 564
keys don't respond 565
pineapple-slicing 566
potato-peeling machine 566
printer won't print 566
quotation marks not found 563
smart quotations 563
takes ages 566
the English 566
weird symbols 564
won't start 566
WordList out of order 567
t-score
formula 540
T-score computing 344
TTR 353
Two word-list analysis 267
type
definition 533
type and token

© 2019 Mike Scott

type and token
definition 533
type/token ratios 353
type-in mode 535
type-in search 106
types of tag 159
typing characters into Concord 528

- U -

undefined tags 253
underscore 146
underscore tags 492
Unicode 527
Unicode codes 538
unicode explained 538
Unicode test 54
unidirectional collocate statistics 207
Unix to Windows 501
unjoin all entries 325
unjoining entries 325
unmarking 325
unreadable 565
unwanted characters 407
updater.exe 22
updating WordSmith 114
updating your version 22
USB drive 22
USB drive folders 539
USB stick 23
user licence 24
user-defined categories 43
saving 182
user-defined processes 128
UTF16 538
UTF8 538
UTF8 versus UTF16 538
utility programs index 386

- V -

valid character
definition 533
value-added annotation 182
version 4 differences 560
Version Checker
overview 11
version checking 25
version date 559
version francaise 544
video
buttons 261
obtaining 263
playing 258
timings 261
view menu 150
Viewer & Aligner 389, 400
adjusting with mouse 392
aligning 388
aligning -- an example 389
aligning the sentences 392
colours 401
dual-text aligning 388
editing 393
index 387
Korean and English aligned text 389
languages 395
moving sentences 392
overview 8
paragraph numbering 395
reading in your plain text 398
sentence joining 400
sentence numbering 395
sentence or paragraph 395
settings 401
splitting 400
technical aspects 401
translation mis-matches 402
troubleshooting 402
unusual sentences 403
viewing options 395
viewing original text file 198
viewing the original text 134

- W -

what is a concordance 196
what you Get settings 253
What's new 4
whole word search 246
why won't it... 563
window management 151
windows

© 2019 Mike Scott

windows
managing 151
Windows file associations 536
Windows XP 548
word
definition 533
word cloud settings
shape 114
word clouds
collocates 219
example 152
word count in MS Word 352
Word documents 552
word patterns 237
word separators
Chinese 534
Japanese 534
Word to .txt 497
word_TAG to <TAG>word 492
WordList
case sensitivity 371
clusters 332
clusters in single-word list 306
coloured tags 374
column visibility 369
comparing word lists 308
comparison display 309
compute keywords 307
create using text file 363
detailed consistency 311
dispersion 354
displaying comparisons 309
find files containing a word 320
finding entries 341
index 305
keys for searching 341
locating entry-types 341
merging 310
minimum & maximum settings 370
n-grams 332
only adjectives 377
overview 7
phrases 306
prefix for tag 374
purpose 305
searching using menu 341
simple consistency 310
sort order 372
sorting problems 567
starting tips 18
summary statistics 359
tags 374
tags as prefix 374
the basic display 380
what you get settings 367
what you see settings 369
WordList Index
clusters 366
computing clusters 556
n-grams 556
omit # 332
relationship settings 366
WordList: altering entries 75
Word's results are different 352
WordSmith controller
Concord settings 253
index settings 366
KeyWords settings 297
WordList settings 367
wordsmith exe file (controller) 5
WordSmith Group 532
WordSmith Tools
installation 22
manual 548
version 559
wordsmith7.ini and networks 25
WSConcGram
aims 510
definition of concgram 511
display 514
exporting concgrams 520
filtering 519
generating concgrams 513
overview 13
settings 512
viewing 514

- X -

X-letter word count 350
XML 543
attributes 189
entities 189
parsing XML 189

© 2019 Mike Scott

XML 543
text handling 189
XML simplification 495

- Y -

Yasumasa Someya 327

- Z -

Z score 342
zapping
filenames recomputed after 154
zip files 561
z-score
formula 540
Z-score computing 344

© 2019 Mike Scott

WordSmith Tools Manual