WordSmith Tools Manual

© 2015 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: February 2015

Publisher

Lexical Analysis Software Ltd.

Special thanks to:

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

## Foreword

<table>
<thead>
<tr>
<th>Part</th>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part I</td>
<td>WordSmith Tools</td>
<td>2</td>
</tr>
<tr>
<td>Part II</td>
<td>Overview</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>Requirements</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>What's new in version 6</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>Controller</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>Concord</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>KeyWords</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>WordList</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>Utilities</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Character Profiler</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>CharGrams</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Choose Languages</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>Corpus Corruption Detector</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>File Utilities</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Splitter</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>File Viewer</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Minimal Pairs</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>Text Converter</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Version Checker</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Viewer and Aligner</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Webgetter</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>WSConcGram</td>
<td>12</td>
</tr>
<tr>
<td>Part III</td>
<td>Getting Started</td>
<td>14</td>
</tr>
<tr>
<td>1</td>
<td>getting started with Concord</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>getting started with KeyWords</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>getting started with WordList</td>
<td>17</td>
</tr>
<tr>
<td>Part IV</td>
<td>Installation and Updating</td>
<td>21</td>
</tr>
<tr>
<td>1</td>
<td>installing WordSmith Tools</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>what your licence allows</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>site licence defaults</td>
<td>23</td>
</tr>
<tr>
<td>4</td>
<td>version checking</td>
<td>24</td>
</tr>
<tr>
<td>Part V</td>
<td>Controller</td>
<td>27</td>
</tr>
<tr>
<td>1</td>
<td>characters and letters</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>accents and other characters</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>wildcards</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>add notes</td>
<td>29</td>
</tr>
</tbody>
</table>

© 2015 Mike Scott
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 adjust settings</td>
<td>29</td>
</tr>
<tr>
<td>4 advanced settings</td>
<td>31</td>
</tr>
<tr>
<td>5 batch processing</td>
<td>39</td>
</tr>
<tr>
<td>6 choosing texts</td>
<td>42</td>
</tr>
<tr>
<td>text formats</td>
<td>42</td>
</tr>
<tr>
<td>the file-choose window</td>
<td>44</td>
</tr>
<tr>
<td>favourite texts</td>
<td>50</td>
</tr>
<tr>
<td>7 choosing files from standard dialogue box</td>
<td>50</td>
</tr>
<tr>
<td>8 class or session instructions</td>
<td>50</td>
</tr>
<tr>
<td>9 colour categories</td>
<td>51</td>
</tr>
<tr>
<td>10 colours</td>
<td>58</td>
</tr>
<tr>
<td>11 column totals</td>
<td>60</td>
</tr>
<tr>
<td>12 compute new column of data</td>
<td>61</td>
</tr>
<tr>
<td>13 copy your results</td>
<td>64</td>
</tr>
<tr>
<td>14 count data frequencies</td>
<td>64</td>
</tr>
<tr>
<td>15 custom processing</td>
<td>65</td>
</tr>
<tr>
<td>16 editing</td>
<td>68</td>
</tr>
<tr>
<td>reduce data to n entries</td>
<td>68</td>
</tr>
<tr>
<td>reverse-delete</td>
<td>69</td>
</tr>
<tr>
<td>delete or restore to end</td>
<td>69</td>
</tr>
<tr>
<td>delete if</td>
<td>69</td>
</tr>
<tr>
<td>editing column headings</td>
<td>69</td>
</tr>
<tr>
<td>editing a list of data</td>
<td>70</td>
</tr>
<tr>
<td>17 find relevant files</td>
<td>73</td>
</tr>
<tr>
<td>18 folder settings</td>
<td>76</td>
</tr>
<tr>
<td>19 fonts</td>
<td>76</td>
</tr>
<tr>
<td>20 main settings</td>
<td>78</td>
</tr>
<tr>
<td>21 language</td>
<td>79</td>
</tr>
<tr>
<td>Overview</td>
<td>81</td>
</tr>
<tr>
<td>Language Chooser</td>
<td>82</td>
</tr>
<tr>
<td>Other Languages</td>
<td>84</td>
</tr>
<tr>
<td>Font</td>
<td>84</td>
</tr>
<tr>
<td>Sort Order</td>
<td>84</td>
</tr>
<tr>
<td>saving your choices</td>
<td>85</td>
</tr>
<tr>
<td>22 layout &amp; format</td>
<td>85</td>
</tr>
<tr>
<td>23 marking entries</td>
<td>90</td>
</tr>
<tr>
<td>24 match words in list</td>
<td>90</td>
</tr>
<tr>
<td>25 previous lists</td>
<td>96</td>
</tr>
<tr>
<td>26 print and print preview</td>
<td>96</td>
</tr>
<tr>
<td>27 quit WordSmith</td>
<td>99</td>
</tr>
<tr>
<td>28 saving</td>
<td>99</td>
</tr>
<tr>
<td>save results</td>
<td>99</td>
</tr>
<tr>
<td>save as text</td>
<td>100</td>
</tr>
<tr>
<td>29 scripting</td>
<td>104</td>
</tr>
</tbody>
</table>
Part VI Tags and Markup 127

1 overview ........................................................................................................ 127
  2 choices in handling tags ............................................................................. 128
  3 custom settings ............................................................................................. 129
  4 tags as selectors .............................................................................................. 130
  5 only if containing... ....................................................................................... 133
  6 part of file: selecting within texts ............................................................... 134
  7 making a tag file .............................................................................................. 137
  8 tag-types .......................................................................................................... 141
  9 start and end of text segments ..................................................................... 142
 10 multimedia tags .............................................................................................. 143
 11 modify source texts ....................................................................................... 145
 12 XML text ......................................................................................................... 148

Part VII Concord 154

1 purpose .............................................................................................................. 154
  2 index ............................................................................................................... 154
  3 what is a concordance? ................................................................................... 155
  4 search-word or phrase ................................................................................... 155
      search word syntax ....................................................................................... 155
      file-based search-words ............................................................................... 157
      search-word and other settings .................................................................... 159
  5 advice .............................................................................................................. 161
  6 blanking .......................................................................................................... 164
  7 Category or Set ................................................................................................. 164
Part VIII KeyWords

1 purpose ........................................................................................................ 223
2 index ............................................................................................................... 223
3 ordinary two word-list analysis ............................................................... 224
4 associate definition .................................................................................... 225
5 associates .................................................................................................... 226
6 choosing files .............................................................................................. 228
7 clumps .......................................................................................................... 230
8 concordance ................................................................................................. 231
9 creating a database ..................................................................................... 231
10 example of key words .............................................................................. 233
11 key key-word definition .......................................................................... 234
12 keyness ........................................................................................................ 234
   p value ........................................................................................................ 234
   key-ness definition .................................................................................... 235
Part IX  WordList

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>thinking about keyness</td>
<td>236</td>
</tr>
<tr>
<td>13 KeyWords database</td>
<td>236</td>
</tr>
<tr>
<td>14 keywords database related clusters</td>
<td>237</td>
</tr>
<tr>
<td>15 KeyWords: advice</td>
<td>237</td>
</tr>
<tr>
<td>16 KeyWords: calculation</td>
<td>238</td>
</tr>
<tr>
<td>17 KeyWords clusters</td>
<td>239</td>
</tr>
<tr>
<td>18 KeyWords: links</td>
<td>240</td>
</tr>
<tr>
<td>19 make a word list from keywords data</td>
<td>242</td>
</tr>
<tr>
<td>20 plot calculation</td>
<td>242</td>
</tr>
<tr>
<td>21 plot display</td>
<td>242</td>
</tr>
<tr>
<td>22 regrouping clumps</td>
<td>244</td>
</tr>
<tr>
<td>23 re-sorting: KeyWords</td>
<td>245</td>
</tr>
<tr>
<td>24 the key words screen</td>
<td>246</td>
</tr>
<tr>
<td>25 WordSmith controller: KeyWords settings</td>
<td>247</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Part IX  WordList</td>
<td>251</td>
</tr>
<tr>
<td>1 purpose</td>
<td>251</td>
</tr>
<tr>
<td>2 index</td>
<td>251</td>
</tr>
<tr>
<td>3 comparing wordlists</td>
<td>252</td>
</tr>
<tr>
<td>4 merging wordlists</td>
<td>253</td>
</tr>
<tr>
<td>5 comparison display</td>
<td>254</td>
</tr>
<tr>
<td>6 consistency analysis (range)</td>
<td>255</td>
</tr>
<tr>
<td>7 compute key words</td>
<td>256</td>
</tr>
<tr>
<td>8 detailed consistency analysis</td>
<td>256</td>
</tr>
<tr>
<td>9 detailed consistency relations</td>
<td>260</td>
</tr>
<tr>
<td>10 find filenames</td>
<td>261</td>
</tr>
<tr>
<td>11 Lemmas (joining words)</td>
<td>262</td>
</tr>
<tr>
<td>what are lemmas and how do we join words?</td>
<td>262</td>
</tr>
<tr>
<td>manual joining</td>
<td>263</td>
</tr>
<tr>
<td>auto-joining lemmas</td>
<td>265</td>
</tr>
<tr>
<td>choosing lemma file</td>
<td>266</td>
</tr>
<tr>
<td>12 WordList Index</td>
<td>267</td>
</tr>
<tr>
<td>what is an Index for?</td>
<td>267</td>
</tr>
<tr>
<td>making a WordList Index</td>
<td>267</td>
</tr>
<tr>
<td>index clusters</td>
<td>268</td>
</tr>
<tr>
<td>join clusters</td>
<td>270</td>
</tr>
<tr>
<td>index lists: viewing</td>
<td>274</td>
</tr>
<tr>
<td>index exporting</td>
<td>275</td>
</tr>
<tr>
<td>13 menu search</td>
<td>279</td>
</tr>
<tr>
<td>14 relationships between words</td>
<td>280</td>
</tr>
<tr>
<td>mutual information and other relations</td>
<td>280</td>
</tr>
<tr>
<td>relationships display</td>
<td>280</td>
</tr>
<tr>
<td>relationships computing</td>
<td>285</td>
</tr>
</tbody>
</table>
Part X Utility Programs

1 Convert Data from Previous Versions .......................................................... 315
   Convert Data from Previous Versions ...................................................... 315
2 WebGetter ...................................................................................................... 315
   overview .................................................................................................... 315
   settings ..................................................................................................... 317
   display ....................................................................................................... 318
   limitations .................................................................................................. 320
3 Corpus Corruption Detector ........................................................................ 321
   Aim ........................................................................................................... 321
   How it works ............................................................................................. 321
4 Minimal Pairs ............................................................................................... 323
   aim ............................................................................................................ 323
   requirements ............................................................................................ 324
   choosing your files .................................................................................. 325
   output ....................................................................................................... 328
   rules and settings .................................................................................... 329
   running the program .............................................................................. 330
5 File Viewer .................................................................................................... 332
   Using File Viewer .................................................................................... 332
6 File Utilities .................................................................................................. 334
   index ........................................................................................................ 334
   Splitter ..................................................................................................... 335
   Splitter: index .......................................................................................... 335
   aim of Splitter ......................................................................................... 335
   Splitter: filenames .................................................................................. 336
   Splitter: wildcards .................................................................................. 337
   join text files ............................................................................................ 338
   compare two files .................................................................................... 339
   file chunker ............................................................................................. 340
   find duplicates ........................................................................................ 340
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>rename</td>
<td>341</td>
</tr>
<tr>
<td>move files to sub-folders</td>
<td>343</td>
</tr>
<tr>
<td>dates and times</td>
<td>344</td>
</tr>
<tr>
<td>find holes in texts</td>
<td>346</td>
</tr>
<tr>
<td><strong>7 Text Converter</strong></td>
<td>346</td>
</tr>
<tr>
<td>purpose</td>
<td>346</td>
</tr>
<tr>
<td>index</td>
<td>347</td>
</tr>
<tr>
<td>settings</td>
<td>347</td>
</tr>
<tr>
<td>Text Converter: copy to</td>
<td>350</td>
</tr>
<tr>
<td>extracting from files</td>
<td>351</td>
</tr>
<tr>
<td>filtering: move if</td>
<td>352</td>
</tr>
<tr>
<td>Convert within the text file</td>
<td>354</td>
</tr>
<tr>
<td>conversion file</td>
<td>355</td>
</tr>
<tr>
<td>syntax</td>
<td>356</td>
</tr>
<tr>
<td>sample conversion file</td>
<td>357</td>
</tr>
<tr>
<td>Convert format of entire text files</td>
<td>358</td>
</tr>
<tr>
<td>Mark-up changes</td>
<td>361</td>
</tr>
<tr>
<td>Word, Excel, PDF</td>
<td>364</td>
</tr>
<tr>
<td>non-Unicode Text</td>
<td>365</td>
</tr>
<tr>
<td>Other changes</td>
<td>366</td>
</tr>
<tr>
<td>Text Converter: converting BNC XML version</td>
<td>367</td>
</tr>
<tr>
<td><strong>8 Viewer and Aligner</strong></td>
<td>372</td>
</tr>
<tr>
<td>purpose</td>
<td>372</td>
</tr>
<tr>
<td>index</td>
<td>373</td>
</tr>
<tr>
<td>aligning with Viewer &amp; Aligner</td>
<td>374</td>
</tr>
<tr>
<td>example of aligning</td>
<td>374</td>
</tr>
<tr>
<td>aligning and moving</td>
<td>377</td>
</tr>
<tr>
<td>editing</td>
<td>377</td>
</tr>
<tr>
<td>languages</td>
<td>378</td>
</tr>
<tr>
<td>numbering sentences &amp; paragraphs</td>
<td>379</td>
</tr>
<tr>
<td>options</td>
<td>379</td>
</tr>
<tr>
<td>reading in a plain text</td>
<td>380</td>
</tr>
<tr>
<td>joining and splitting</td>
<td>382</td>
</tr>
<tr>
<td>settings</td>
<td>383</td>
</tr>
<tr>
<td>technical aspects</td>
<td>383</td>
</tr>
<tr>
<td>translation mis-matches</td>
<td>383</td>
</tr>
<tr>
<td>troubleshooting</td>
<td>384</td>
</tr>
<tr>
<td>unusual lines</td>
<td>384</td>
</tr>
<tr>
<td><strong>9 WSCconcGram</strong></td>
<td>385</td>
</tr>
<tr>
<td>aims</td>
<td>385</td>
</tr>
<tr>
<td>definition of a concgram</td>
<td>386</td>
</tr>
<tr>
<td>settings</td>
<td>387</td>
</tr>
<tr>
<td>generating concgrams</td>
<td>387</td>
</tr>
<tr>
<td>viewing concgrams</td>
<td>389</td>
</tr>
<tr>
<td>filtering concgrams</td>
<td>394</td>
</tr>
<tr>
<td>exporting concgrams</td>
<td>396</td>
</tr>
<tr>
<td><strong>10 CharacterProfiler</strong></td>
<td>397</td>
</tr>
<tr>
<td>purpose</td>
<td>397</td>
</tr>
<tr>
<td>profiling characters</td>
<td>397</td>
</tr>
<tr>
<td>profiling settings</td>
<td>400</td>
</tr>
<tr>
<td><strong>11 Chargrams</strong></td>
<td>401</td>
</tr>
<tr>
<td>purpose</td>
<td>401</td>
</tr>
</tbody>
</table>
Part XI Reference

1 acknowledgements ................................................................. 408
2 API ...................................................................................... 408
3 bibliography ......................................................................... 409
4 bugs .................................................................................... 409
5 change language ................................................................... 411
6 Character Sets ....................................................................... 411
   overview .............................................................................. 411
   accents & symbols ............................................................. 412
7 clipboard .............................................................................. 414
8 contact addresses ................................................................... 417
9 date format ............................................................................ 417
10 Definitions ............................................................................ 417
   definitions ............................................................................ 417
   word separators ................................................................... 419
11 demonstration version ........................................................... 419
12 drag and drop ....................................................................... 419
13 edit v. type-in mode ............................................................. 420
14 file extensions ...................................................................... 420
15 finding source texts .............................................................. 422
16 flavours of Unicode .............................................................. 422
17 folders/directories ................................................................. 423
18 formulae .............................................................................. 425
19 history list ............................................................................ 427
20 HTML, SGML and XML ......................................................... 427
21 hyphens .............................................................................. 427
22 international versions ............................................................ 428
23 limitations ............................................................................ 429
   tool-specific limitations ...................................................... 429
24 links between tools ............................................................... 430
25 keyboard shortcuts ............................................................... 431
26 machine requirements ........................................................... 432
27 manual for WordSmith Tools ................................................ 432
28 menu and button options ...................................................... 433
29 MS Word documents ............................................................. 436
30 never used WordSmith before ............................................... 438
31 numbers .............................................................................. 438
32 plot dispersion value ........................................................... 438
### Part XII Troubleshooting

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 list of FAQs</td>
<td>448</td>
</tr>
<tr>
<td>2 apostrophes not found</td>
<td>448</td>
</tr>
<tr>
<td>3 column spacing</td>
<td>448</td>
</tr>
<tr>
<td>4 Concord tags problem</td>
<td>448</td>
</tr>
<tr>
<td>5 Concord/WordList mismatch</td>
<td>449</td>
</tr>
<tr>
<td>6 crashed</td>
<td>449</td>
</tr>
<tr>
<td>7 demo limit</td>
<td>449</td>
</tr>
<tr>
<td>8 funny symbols</td>
<td>449</td>
</tr>
<tr>
<td>9 illegible colours</td>
<td>450</td>
</tr>
<tr>
<td>10 keys don’t respond</td>
<td>450</td>
</tr>
<tr>
<td>11 pineapple-slicing</td>
<td>451</td>
</tr>
<tr>
<td>12 printer didn’t print</td>
<td>451</td>
</tr>
<tr>
<td>13 too slow</td>
<td>451</td>
</tr>
<tr>
<td>14 won’t start</td>
<td>451</td>
</tr>
<tr>
<td>15 word list out of order</td>
<td>452</td>
</tr>
</tbody>
</table>

### Part XIII Error Messages

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 list of error messages</td>
<td>454</td>
</tr>
<tr>
<td>2 .ini file not found</td>
<td>455</td>
</tr>
<tr>
<td>3 administrator rights</td>
<td>455</td>
</tr>
<tr>
<td>4 base list error</td>
<td>456</td>
</tr>
<tr>
<td>5 can only save words as ASCII</td>
<td>456</td>
</tr>
<tr>
<td>6 can’t call other tool</td>
<td>456</td>
</tr>
<tr>
<td>7 can’t make folder as that’s an existing filename</td>
<td>456</td>
</tr>
<tr>
<td>8 can’t compute key words as languages differ</td>
<td>456</td>
</tr>
<tr>
<td>9 can’t merge list with itself!</td>
<td>456</td>
</tr>
<tr>
<td>10 can’t read file</td>
<td>456</td>
</tr>
</tbody>
</table>
11 character set reset to <x> to suit <language> .................................................. 457
12 concordance file is faulty ................................................................................. 457
13 concordance stop list file not found ................................................................. 457
14 confirmation messages: okay to re-read ......................................................... 457
15 conversion file not found .................................................................................. 457
16 destination folder not found ............................................................................. 458
17 disk problem -- file not saved ......................................................................... 458
18 dispersions go with concordances .................................................................... 458
19 drive not valid .................................................................................................... 458
20 failed to access Internet ..................................................................................... 458
21 failed to create new folder name ....................................................................... 458
22 failed to read file ............................................................................................... 458
23 failed to save file ............................................................................................... 459
24 file access denied ............................................................................................. 459
25 file contains none of the tags specified ............................................................ 459
26 file has "holes" ................................................................................................... 459
27 file not found ..................................................................................................... 459
28 filenames must differ! ....................................................................................... 460
29 folder is read-only ............................................................................................ 460
30 for use on X machine only ................................................................................ 460
31 form incomplete ................................................................................................ 460
32 full drive & folder name needed ....................................................................... 460
33 function not working properly yet ...................................................................... 460
34 invalid concordance file ................................................................................... 460
35 invalid file name ................................................................................................ 460
36 invalid KeyWords database file ......................................................................... 461
37 invalid KeyWords calculation ........................................................................... 461
38 invalid WordList comparison file ...................................................................... 461
39 invalid WordList file ......................................................................................... 461
40 joining limit reached ........................................................................................ 461
41 KeyWords database file is faulty ....................................................................... 462
42 KeyWords file is faulty ....................................................................................... 462
43 limit of file-based search-words reached ......................................................... 462
44 links between Tools disrupted ......................................................................... 462
45 match list details not specified .......................................................................... 462
46 must be a number .............................................................................................. 462
47 mutual information incompatible ..................................................................... 463
48 network registration used elsewhere ............................................................... 463
49 no access to text file - in use elsewhere? ......................................................... 463
WordSmith Tools Manual

50 no associates found ................................................................. 463
51 no clumps identified ............................................................... 463
52 no clusters found ................................................................. 463
53 no collocates found ............................................................... 463
54 no concordance entries ......................................................... 464
55 no concordance stop list words .............................................. 464
56 no deleted lines to zap ......................................................... 464
57 no entries in KeyWords database .......................................... 464
58 no fonts available ............................................................... 464
59 no key words found ............................................................ 464
60 no key words to plot ........................................................... 465
61 no KeyWords stop list words ................................................. 465
62 no lemma list words ............................................................ 465
63 no match list words ............................................................. 465
64 no room for computed variable ............................................. 465
65 no statistics available ........................................................ 465
66 no stop list words ............................................................... 465
67 no such file(s) found ............................................................ 465
68 no tag list words ................................................................. 466
69 no word lists selected ........................................................ 466
70 not a valid number ............................................................. 466
71 not a WordSmith file ........................................................... 466
72 not a current WordSmith file ............................................... 466
73 nothing activated ............................................................... 466
74 Only X% of words found in reference corpus ......................... 467
75 original text file needed but not found .................................. 467
76 printer needed ................................................................. 467
77 registration code in wrong format ........................................... 467
78 registration is not correct ..................................................... 468
79 short of memory ................................................................. 468
80 source folder file(s) not found ............................................... 468
81 stop list file not found ........................................................ 468
82 stop list file not read ........................................................... 468
83 tag file not found ............................................................... 468
84 tag file not read ................................................................. 468
85 this function is not yet ready ................................................ 469
86 this is a demo version ......................................................... 469
87 this program needs Windows XP or greater .......................... 469
88 to stop getting this message ................................................ 469
<table>
<thead>
<tr>
<th>Page</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
<td>too many requests to ignore matching clumps</td>
</tr>
<tr>
<td>90</td>
<td>too many sentences</td>
</tr>
<tr>
<td>91</td>
<td>truncating at xx words -- tag list file has more</td>
</tr>
<tr>
<td>92</td>
<td>two files needed</td>
</tr>
<tr>
<td>93</td>
<td>unable to merge Keywords Databases</td>
</tr>
<tr>
<td>94</td>
<td>why did my search fail?</td>
</tr>
<tr>
<td>95</td>
<td>word list file is faulty</td>
</tr>
<tr>
<td>96</td>
<td>word list file not found</td>
</tr>
<tr>
<td>97</td>
<td>WordList comparison file is faulty</td>
</tr>
<tr>
<td>98</td>
<td>WordSmith Tools already running</td>
</tr>
<tr>
<td>99</td>
<td>WordSmith Tools expired</td>
</tr>
<tr>
<td>100</td>
<td>WordSmith version mis-match</td>
</tr>
<tr>
<td>101</td>
<td>XX days left</td>
</tr>
</tbody>
</table>

Index 471
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**, **Concord**, or **KeyWords**.

Version:6.0

© 2015 Mike Scott
Overview

Section II
2 Overview

2.1 Requirements

WordSmith Tools requires

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

2.2 What's new in version 6

WordSmith is organic software!

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

New features:

- Move files to sub-folders
- Skins
- Word Clouds
- Date handling & Time-lines
- .docx files
- Scripting
- Colour categories
- Phrase frames
- Collocate following
- Chargsms

2.3 Controller

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

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

You can minimise it, if you feel the screen is getting cluttered.
Concord is a program which makes a concordance 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 for later use, edited, printed, copied to your word-processor, or saved as text files.

See also: Concord Help Contents Page, The buttons

### 2.5 KeyWords

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

The distribution of the key words can be plotted.

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

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

See also: KeyWords Help Contents Page, The buttons

### 2.6 WordList

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

See also: WordList Help Contents Page

### 2.7 Utilities

#### 2.7.1 Character Profiler
A tool to help find out which characters or chargrams 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

2.7.2 CharGrams

A tool to help find out which chargrams (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

2.7.3 Choose Languages

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

See also: Choose Language Tool

2.7.4 Corpus Corruption Detector

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

2.7.5  **File Utilities**

Utilities to

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

2.7.5.1  **Splitter**

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

See also: Splitter Help Contents Page

2.7.6  **File Viewer**

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

See also: File Viewer Index

2.7.7  **Minimal Pairs**

a program to find typos and minimally-differing pairs of words.
See also : aim, requirements, choosing your files, output, rules and settings, running the program.

2.7.8 Text Converter

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

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

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

See also: Text Converter Help Contents Page

2.7.9 Version Checker

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

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

2.7.10 Viewer and Aligner

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

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

See also: Viewer & Aligner Help Contents Page
2.7.11 Webgetter

A tool to gather text from the Internet.

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

See also: A fuller overview, Settings, Display, Limitations

2.7.12 WSConcGram

a tool for generating concgrams

See also: Aims of WSConcGram, Running WSConcGram
Getting Started

Section III
3 Getting Started

3.1 getting started with Concord

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

In the main WordSmith Tools window (the one with WordSmith Tools Controller 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 or change your choice, and make a new concordance, looking somewhat like this:

(If you only see the window with Concord in its caption, choose File | New and the Getting Started window will open up.)
If you have never used WordSmith before you will find a text has been selected for you automatically to help you get started. You will need to specify a Search-Word or phrase and then press OK.

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

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

If you want to alter other settings, press Advanced, but you can probably leave the default settings as they are.
Concord now searches through your text(s) looking for the search word or Tag.

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

See also: Concord Help Contents.

3.2 getting started with KeyWords

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

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

Choose File | New

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

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

You'll find that KeyWords starts processing your file and a progress window in the main Controller shows a bar indicating how it's getting on. After KeyWords has finished, it will show you a list of the key words. The ones at the top are more "key" than those further down.
Don't forget to save the results (press Ctrl+F2) if you want to keep the keyword list for another time.

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

### 3.3 getting started with WordList

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

I suggest you start by trying the WordList program. In the main WordSmith Tools window (the one with WordSmith Tools Controller 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 or change your choice, and make a new word list.
If you have never used WordSmith before you will find a text has been selected for you automatically to help you get started.

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

You'll find that WordList starts processing your file(s) and a progress window in the main Controller shows a bar indicating how it's getting on. After WordList has finished making the list, you will see some windows showing the words from your text file in alphabetical order and in frequency order, statistics, filenames, notes.
Don't forget to **save the results** (press Ctrl+F2 or ![button]) if you want to keep the word list for another time.

See also: [WordList Help Contents](#).
Installation and Updating

Section IV
4 Installation and Updating

4.1 installing WordSmith Tools

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

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

--- Single User Licence

Everything must correspond exactly to what you were given when you purchased. Paste in your Name as specified in your purchase email or screen and (if there are any in the registration) Other Details, and paste in the code.

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

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

Install to a removable drive

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

## Site Licence

Follow the instructions at


## 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 [http://lexically.net/wordsmith/purchasing.htm](http://lexically.net/wordsmith/purchasing.htm) for details of suppliers.

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

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

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

### 4.2 what your licence allows

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

**SINGLE USER LICENCES**

*Think of these as a licence for a person.*
You can install the product on a machine at your office and a machine at home. You may
yourself use both copies of the product, but only one at a time.
You cannot install the product on two machines, and then use both of those copies at the same
time, or allow anyone else to use your copy of the product on the second machine. For
instance, you cannot purchase one copy of the product and allow a friend or family member to
use the product on the other machine.
You may not, at any time, allow another user to install your copy of the product for his/her own
use.

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

The full licence text is at \wsmith6\user_licence.txt.

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

the wordsmith6.ini file

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

The following terms are used
prohibited drives
limited folder
instructions folder
network-read/write folder

and an example would be

[NETWORK]
network-read/write folder=m:\Documents\wsmith6
(prohibitted drives=xyz

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

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

See also: What your licence allows, Class Instructions

4.4 version checking

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

![version checking](image)

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, version updating.
5 Controller

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

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

5.1.1 accents and other characters

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

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

5.1.2 wildcards

Many WordSmith functions allow you a choice of wildcards:

<table>
<thead>
<tr>
<th>symbol</th>
<th>meaning</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>*</td>
<td>disregard the end of the word, disregard a whole word</td>
<td>tele*</td>
</tr>
<tr>
<td></td>
<td></td>
<td>*ness</td>
</tr>
<tr>
<td></td>
<td></td>
<td><em>happi</em></td>
</tr>
</tbody>
</table>
book * hotel

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

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

5.2 add notes

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

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, you will probably want to list these and save the information for later use.

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

5.3 adjust settings

There are a number of Settings windows in the Controller. You will see tabs accessing them at the left in the main Controller window.
Choose and save settings concerning:

- **font**
- **colours**
- **folders**
- **tags**
- **general settings**
- **match-lists**
- **stop lists**
- **lemma lists**
- **text and language settings**
- **Concord Settings**
- **KeyWords settings**
- **WordList settings**
- **advanced user specific settings**
- **index file settings**
5.4 advanced settings

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

![Advanced Settings]

and open up a whole new set of options

![Advanced Settings]

- Tags & Markup
- Lists
- Index
- Scripts
Help, logging

Help system access
On a network, it is commonly the case that Microsoft protects users to such an extent that the usual .CHM help files show only their table of contents but no details. Here you can set the WordSmith help to access the local CHM file, a local set of HTML files 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\wsmith6 <Enter> | wordsmith6 /log <Enter>
(or wordsmith6 /log C:\temp\WSLog.txt to force use of C:\temp\WSLog.txt. If you do that, make sure the folder exists first.)

See also: emailed error reports
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

Advanced section (menus, clipboard, deadkeys etc.)

Customising menus

You can re-assign new shortcuts (such as Alt+F3, Ctrl+O) to the menu items 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.
You will see a list of menu options at the left, and can add to (or remove from) the list on the right by selecting one on the left and pressing the buttons in the middle, or by dragging it to the right. To re-order the choices, press the up or down arrow. In the screenshot I've added "Concordance" as I usually want to generate concordances from word-lists and key word lists.

Whatever is in your popup menu will also appear in the Toolbars.

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

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

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

*prompt to save (in general)*: reminds you to save every time new data results are computed or reorganised.

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

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

**Clipboard**

Here you may choose defaults for copying.
The number of characters only applies when copying as editable text. See also: clipboard

**User .dll**

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

**Language Input**

*Deadkeys* are used to help type accented characters with some keyboards. The language input tab lets you alter the deadkeys to suit your keyboard and if necessary force WordSmith to use the keyboard layout of your choice whenever WordSmith starts up.
Here the user's Windows has four keyboard layouts installed. To type in Maori, you might choose to select Maori, and change a couple of deadkeys. At present, as the list shows, pressing ` then A gives À, but users of Maori usually prefer that combination to give ā.

To change these settings,
1. select the line

2. edit the box below:

(you can drag the character you need from the character window)

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

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

- **Convert in place**
- **Convert into temp folder**
- **Without confirmation**
- **Confirm each one**
- **Convert from UTF8**
- **reserved**

See also: menu and button options

5.5 batch processing

The point of it...

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

If you have selected more than one text file you can ask WordList, Concord and KeyWords to process as a batch.
Folder where they end up
The name suggested is today’s date. Edit it if you like. Whatever you choose will get created when the batch process starts.

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

```
c:\wsmith\wordlist\05_07_19_12_00,
```

they will end up like this:

```
c:\wsmith\wordlist\05_07_19_12_00\fred1.lst
```

```
c:\wsmith\wordlist\05_07_19_12_00\jim2.lst
```

```
.. c:\wsmith\wordlist\05_07_19_12_00\mary512.lst
```

then

```
c:\wsmith\wordlist\05_07_19_12_00\1\joanna513.lst
```

etc.

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

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

If you zip them, you will get

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

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

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

But what if your text files are structured like this:

```
\..\BNC
\..\BNC\written
```
The One file per folder, individual zip files 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 zip files 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**

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

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

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

You specify

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

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

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

If you did not check "zip them into 1 .zip file", you will find them under C:\WSMITH\WORDLIST
\\ELEPHANT\\0.
If you did check "zip them into 1 .zip file", there is now a C:\WSMITH\WORDLIST\\ELEPHANT.ZIP file
which contains all your results. (The 1,200 .LST files created will have been erased but the .ZIP file
contains all your lists.)

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

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

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

See also: batch scripts

5.6 choosing texts

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

You need text in a suitable format.

5.6.1 text formats

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

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

**Why not .PDF files?**

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

**Why not .DOC files?**

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

How to get here

This function is accessed from the File menu in the **Controller** and the Settings menu or New menu item ( ) in the various Tools.
The two main areas at left and right are
- files to choose from (at left)
- files already selected (at right)

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

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

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.

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.

---

**File Types**

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

---

**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.

---

**Test for Unicode**

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**

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

---

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

(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.)

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

There is also a utility in the File Utilities[448] which can parse text files to generate dates using your own syntax, preparing a text file like this to read in.

You can save the dates and files as favourites[569] so as to re-use this information as often as you like.

See also: using text dates[127]

---

### 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.

**View**

Allows you to browse within the currently selected file so as to check whether to include it. Any accented characters (e.g. æ, é) or currency symbols such as £, ¥, ¢, and tags will appear according to current Text Characteristics settings[123]. You can change these while viewing the file.

**Sorting**

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

**View in Notepad**

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

**Get from Internet**

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

**Zip files**

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

**Save List**

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

**save favourites**

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

Saves a list of text files whose status is either unknown or known to meet your requirements when selecting files by their contents, 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](#) [file dates](#)

5.7 choosing files from standard dialogue box

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

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

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

5.8 class or session instructions

When WordSmith is run in a training session, you may want to make certain instructions available to your trainees.
To do this, all you need to do is ensure there is a file called `teacher.rtf` in your main `\wsmith6` folder where the WordSmith programs are or in the "instructions folder" explained under [site licence defaults](#). 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](#)

### 5.9 colour categories

The point of it ...

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

The idea is to let you re-sort 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` 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 data column (above the user chose the File column) and a condition (here X ends with Y and .txt below which will mean 'search the file column seeking any where the File ends in .txt'). Then choose a colour (here colour 97 was chosen) and then press Add a search. Finally, press Find.

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

A more complex example:
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.
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.

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.
Once sorted, the data can be saved as before.

**What if I already have a Set classification?**

Here is a concordance where the exclamation O or Ah had already been identified and marked in the Set column.
As the Set column is already in use, classifying further by colour will take second priority to the existing forms typed in. So in this case:

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

but clicking the Set column gives priority to the exclamation typed in.
In this case the Os follow the Ah's, and the coloured Ah's follow the uncoloured ones.

**Removing the colours?**

Use the Clear colours button.

**What if more than one condition is met?**

If you colour words ending NESS blue, and also colour words starting UN yellow, any word meeting both conditions will get a mixture of the two colours as shown here:
See also: setting categories by typing, colour categories for concordances

5.10 colours

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

- **plain text**: this is the default colour
- **highlighted text**: as above when selected
- **tags**: mark-up
- **search word**: concordance search word; words in (key) word lists
- **main sort word**: indicates first sort preference; used for % data in (key) word lists
- **second sort word**: indicates first tie-breaker sort colour
- **context word**: context word
- **deleted words**: any line of deleted data
- **not numbered line**: any line which has not been user-sorted
- **search word**: concordance search word when selected
- **highlighted**: first sort when selected
- **main sort word**: first tie-breaker sort when selected
- **highlighted**: context word when selected
- **highlighted**: most frequent collocate or detailed consistency word, p value
- **viewing texts**: in the text viewer
- **lemma colour**: colour of lemmas shown in lemma window
- **word-cloud shape**: see word clouds section below
- **word-cloud window**:
Overall colour scheme
This allows a range of colour scheme choices, which will affect the colours of all WordSmith windows.

List colours
To alter colours, first click on the wording you wish to change (you'll see a difference in the left margin: here search word has been chosen), then click on a colour in the colour box. The Foreground and Background radio buttons determine whether you're changing foreground or background colours. You can press the Reset button if you want to revert to standard defaults. The same colours, or equivalent shades of grey, will appear in printouts, or you can set the printer to black and white, in which case any column not using "plain text" colour will appear in italics (or bold or underlined if you have already set the column to italics).

The Reset button lets you restore colours to factory defaults.

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

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

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

### 5.11 column totals

**The point of it...**

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

**How to do it**

With a word-list, concordance or key-words list visible, choose the menu item View / Column Totals to switch column totals on or off.
Here we see column totals on a detailed consistency list based on Shakespeare's plays. The list itself is sorted by the Texts column: the top items are found in all 35 of the plays used for the list. In the case of Anthony and Cleopatra, A represents 1.28% of the words in that column, that is 1.28% of the words of the play Anthony and Cleopatra. In the case of ACT this is the highest percentage in its row (this word is used more in percentage terms in that play than in the others).

See also: save as Excel

5.12 compute new column of data

The point of it...

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

<table>
<thead>
<tr>
<th>you type</th>
<th>Result -- for each row in your data, the new column will contain:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rel(2) ÷ 5</td>
<td>the data from column 2 of the same row, divided by 5</td>
</tr>
<tr>
<td>RelC(2)</td>
<td>the data from column 2 of the same row, added to a running total</td>
</tr>
<tr>
<td>Rel(3) + (Rel(2) ÷ 5)</td>
<td>the data from column 2 of the same row, divided by 5, added to the data from column 3 of the same row</td>
</tr>
<tr>
<td>Abs(2;1) ÷ 5</td>
<td>the data from column 2 of row 1, divided by 5. (This example is just to illustrate; it would be silly as it would give the exact same result in every row.)</td>
</tr>
<tr>
<td>Rel(2) ÷ Abs(2;1) × 100</td>
<td>the data from column 2 of the same row divided by column 2 of 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.</td>
</tr>
</tbody>
</table>
You can format (or even delete) any variables computed in this way: see layout.

See also: count data frequencies, column totals.

5.13 copy your results

The quickest and easiest method of copying your data e.g. into your word processor is to select with the cursor arrows and then press Ctrl+C. This puts it into the clipboard (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
- save as data (not the same as saving as text: this is saving so you can access your data again another day)

See also: saving, printing, clipboard.

5.14 count data frequencies

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

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

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

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

Cumulative Column
This adds up values from another column of data. See the example from WordList.
See also: distinguishing consequence from consequences, frequencies of suffixes in a word list, compute new column of data.

5.15 custom processing

This feature -- which, like API, 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:

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

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

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

a) nil (if you simply do not want the original word in your list)
b) the same PCHAR/PWIDECHAR if it is not to be changed at all
c) a replacement form
Here's an example where the source text was

**Today is Easter Day.**

![WordList](image)

### Source code

The source code for the .dll in Delphi is this

```delphi
library WS5WordSmithCustomDLL;
uses
  Windows, SysUtils;
{
  This example uses a very straightforward Windows routine for comparing
  strings, CompareStringA and CompareStringW which are in a Windows .dll.

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

  In this example, EASTER gets changed to CHRISTMAS.
}

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

else begin
  if CompareStringA(
    language_identifier,
    NORM_IGNORECASE,
    PAnsiChar(original), -1,
    PAnsiChar('EASTER'), -1) - 2 = 0
  then
    Result := pAnsichar('CHRISTMAS');
  end;
end;

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

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

{ This routine exports each concordance line together with
  the filename it was found in
  a number stating how many bytes into the source text file the entry was found
  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 HAPINESS this would be 9)
  This information is saved in Unicode appended to your results_filename }

function HandleConcordanceLine
(source_line : pointer;
  hit_pos_in_characters,
  hit_length : integer;
  byte_position_in_file,
  language_id : DWORD;
  is_Unicode : WordBool;
  source_text_filename,
  results_filename : pwidechar) : pointer; stdcall;
begin
  function extrasA : ansistring;
begin
    Result := #9+ ansistring(widestring(pwidechar(source_text_filename)))+
    #9+ ansistring(IntToStr(byte_position_in_file))+
    #9+ ansistring(IntToStr(hit_pos_in_characters))+
    #9+ ansistring(IntToStr(hit_length));
  end;

  function extrasW : widestring;
begin
    Result := #9+ widestring(pwidechar(source_text_filename)) +
    #9+ IntToStr(byte_position_in_file) +
const
  bm: char = widechar($FEFF);
var f : File of widechar;
output_string : widestring;
begin
  Result := source_line;
  if length(results_filename)>0 then try
    AssignFile(f,results_filename);
    if FileExists(results_filename) then begin
      Reset(f);
      Seek(f, FileSize(f));
    end else begin
      Rewrite(f);
      Write(f, bm);
    end;
    if is_Unicode then
      output_string := pwidechar(source_line)+extrasW
    else
      output_string := pAnsichar(source_line)+widestring(extrasA);
  if length(output_string) > 0 then
    BlockWrite(f, output_string[1], length(output_string));
  CloseFile(f);
  except
    end;
  end;
exports
  ConcordChangeWord,
  KeyWordsChangeWord,
  WordlistChangeWord,
  HandleConcordanceLine;
begin
end.

See also: API[408], custom settings[125]

5.16 editing
5.16.1 reduce data to n entries

With a very large word-list, concordance etc., you may wish to reduce it randomly (eg. for sampling). This menu option (Edit | Deleting | Reduce to N) allows you to specify how many entries you want to have in the list. If you reduce the data, entries will be randomly zapped 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[125], editing a list of data[70].
5.16.2 reverse-delete

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

See also: Delete or Restore to End

5.16.3 delete or restore to end

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

See also: Reverse-delete, Reduce to N entries

5.16.4 delete if

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

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

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

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

5.16.5 editing column headings

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

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Freq.</th>
<th>%</th>
<th>Texts</th>
<th>%</th>
<th>Lemmas</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>THE</td>
<td>57861</td>
<td>6.04</td>
<td>4050</td>
<td>98.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>OF</td>
<td>9233</td>
<td>3.05</td>
<td>4040</td>
<td>98.65</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>AND</td>
<td>384</td>
<td>2.52</td>
<td>4060</td>
<td>98.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>ID</td>
<td>55903</td>
<td>2.50</td>
<td>1019</td>
<td>98.66</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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.

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

5.16.6  editing a list of data

With a word list on screen, you might see something like this.
In the status bar at the bottom,

```
AA
```

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

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

If you right-click the second cell you will see

Find
Set
Edit

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

Choose Edit.

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

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

In this case we want to correct AACUTE, which should be Á.
If you now type Á, you will immediately see the result in the window:

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

After editing you may want to re-sort, and if you have changed a word such as AAAAAAGH to a pre-existing word such as AAGH, to join the two entries.

See also: joining entries, finding source files.
5.17  find relevant files

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

How to do it
This function can be reached in any window of data which contains the $F$ menu option, e.g. a word-list or a key words 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 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. 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 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*. 
5.18 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 fonts

Found by choosing Settings / Font in all Tools or via Language Settings in the Controller. Enables you to choose a preferred Windows font and point size for the display windows and printing in all the WordSmith Tools suite. Note that each language can have its own different default font.
If you have data visible in any Tool, the font will automatically change; if you don’t want any specific windows of data to change, because you want different font sizes or different character sets in different windows, minimise these first.

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

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

Found in Main settings in the WordSmith Tools Controller.

**Startup**
*Restore last work* will bring back the last word-list, concordance or key-words list when you start WordSmith.
*Show Help file* will call up the Help file automatically when you start WordSmith.
*Associate/clear file extensions* will teach Windows to use (or not to use) Concord, WordList, KeyWords etc. to open the relevant files made by WordSmith.

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

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

---
**System**

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

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

---
**5.21 language**

**The point of it ...**

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

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

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

**Hyphens separate words, Numbers, Characters within word**

**Font**

**Text Format**

**How more languages are added**

Press the Edit Languages button.

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

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

**How to get here**
The Language Chooser is accessed from the main WordSmith Controller menu: Settings / Main Settings / Text and Languages / Other Languages.

What you will see may look like this:
9 languages have been chosen already.
At the bottom you will see what the fonts on your system are for the current language selected.

See also: Language, Font, Sort Order, Other Languages, saving your choices, Changing the language of saved data.

5.21.2 Language Chooser

How to get here
The Language Chooser is accessed from the main WordSmith Controller:
What it does

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

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

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

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

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

Some languages do not mark word-separators.

See also : Other Languages, saving your choices

5.21.3 Other Languages

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

See also : Language, Font, Sort Order, saving your choices

5.21.4 Font

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

Click here for more on Unicode.

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

5.21.5 Sort Order

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

The display
You will see 2 windows below “Resort” -- the one at the left contains some words in various languages; you can add your own. If your keyboard won’t let you type them in, paste from your own collection of texts.

The one at the right shows how these words get sorted according to the language you have selected.

See also: Language, Font, Other Languages, saving your choices

5.21.6 saving your choices

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

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

5.22 layout & format

With any list open, right-click or choose View | Layout to choose your preferred display formats for each column of data.
Layout or Add data?
The Layout tab gives you a chance to format the layout of your data. Add a column of data lets you compute a new variable.

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

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

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

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

Screen Width
in your preferred units (cm. or inches).
Here 3 of the headings have been activated (by clicking) so that settings can be changed so as to get them all the same width.
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.,

<table>
<thead>
<tr>
<th>Word</th>
<th>%</th>
<th>0.01</th>
<th>0.02</th>
<th>0.05</th>
<th>0.1</th>
<th>0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>WALL</td>
<td>28</td>
<td>0.17</td>
<td>67</td>
<td>108</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>FAIRY</td>
<td>24</td>
<td>0.15</td>
<td>41</td>
<td>105</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>MOON</td>
<td>29</td>
<td>0.16</td>
<td>130</td>
<td>0.02</td>
<td>82</td>
<td></td>
</tr>
<tr>
<td>LOVE</td>
<td>106</td>
<td>0.65</td>
<td>1,948</td>
<td>0.24</td>
<td>77</td>
<td></td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th></th>
<th>WALL</th>
<th>28</th>
<th>0.17</th>
<th>67</th>
<th>0.01</th>
<th>1080</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>FAIRY</td>
<td>24</td>
<td>0.15</td>
<td>41</td>
<td>0.01</td>
<td>1058</td>
</tr>
<tr>
<td></td>
<td>MOON</td>
<td>29</td>
<td>0.13</td>
<td>130</td>
<td>0.02</td>
<td>824</td>
</tr>
<tr>
<td></td>
<td>LOVE</td>
<td>106</td>
<td>0.65</td>
<td>1.948</td>
<td>0.24</td>
<td>774</td>
</tr>
<tr>
<td></td>
<td>LION</td>
<td>24</td>
<td>0.15</td>
<td>88</td>
<td>0.01</td>
<td>765</td>
</tr>
</tbody>
</table>

**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 the concordance in the usual way it will remember your settings anyway. But the next time you make a concordance, you'll get the WordSmith default layout. If you choose this Save, the next time you make a concordance, it will look like the current one.

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

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

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

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>test 679</th>
<th>test 680</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AND</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>EIGHT</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>EIGHTY</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>FIFTY</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>FIVE</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>FORTY</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>FOUR</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>HELLO</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>HUNDRED</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>LINE</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

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

Similarly a statistics list allows the text file-names column to be frozen:
5.23 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.

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

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.

5.24 match words in list

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

**How to do it: WordList example**

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

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

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 \textit{Load} to load its words. You must of course decide what is to be done with any matching entries.

**Text File or Template**

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

To use a match list in a file, you first prepare a file, using \textit{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**

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

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

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

I answered No so you could see this result:
In the screenshot below, the action was *find matches & mark them*, and the match-list contained archaic forms like *thou, thee, thy*. 
The marking can be removed using a menu option or by re-running the match-list function with remove match marking as the action.

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

See also: Comparing Word-lists, Comparing Versions, Stop Lists, Lemma Matching
5.25 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.

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

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

5.26 print and print preview

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

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

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

Next (➡️) & Last (⬅️) Page
Takes you forward or back a page.

Portrait (↕️) or Landscape (↔️)?
Sets printing to the page shape you want.

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

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

**Some columns of data not shown**

A case like this showing nothing but the line numbers

<table>
<thead>
<tr>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>

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 any unwanted ones, or else set the print surface to landscape.

See also: [Printer Settings](#)

### 5.27 quit WordSmith

Alt+X is the hot key.

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

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

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

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

### 5.28 saving

#### 5.28.1 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, 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 or by copying the data to the clipboard.

**save part of the data only**

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

**avoid the save prompts?**

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

### 5.28.2 save as text

**The point of it...**

Save as Text means save your data as a plain text file (as opposed to the WordSmith format for retrieving the data another day). It is usually quicker to copy selected text into the clipboard, 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.

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

**How to do it**

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

Options include:

- **header**: words you want to save at the start of the data (leave blank if 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 (X) save will look something like this:

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

In the case of a concordance line, saving as text will save as many "characters in ‘save as text’" as
you have set (adjustable in the Controller Concord Settings). 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.

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

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

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

At the top you get the raw data, which you can use Excel to create a graphic with.
If you want access to the details of the plot, choose save as text. The results will look like this:

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

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

See also: Excel Files in batch processing
5.29 scripting

Scripts

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

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

```bash
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 subfolders) and saved it.
keywords refcorpus="j:\temp\BNC.lst" wordlist="c:\temp\shakespeare.lst" output="j:\temp\shakespeare.kws"
made a key words list of that word list compared with a BNC word list and saved it.

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

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

If 1_at_a_time is true, a word-list will export separate results text file by text file.
If 1_at_a_time is true, Concord will read search words from a text file and save summary results:
concord corpus="x:\text\dickens\hard_times.txt" node="c:\temp\sws.txt" output="c:\temp\outputs.txt" 1_at_a_time="true"
produced this in c:\temp\outputs.txt:
x:\text\dickens\hard_times.txt
hard 50
soft 3
mean 54
empty 9
fred 0
book 13
north* 4
south* 2

collocate scripts
It is also possible to run a script requesting the collocates of each word in a word-list. This syntax
wordlist collocates of "c:\temp\shakespeare.lst" output="c:\temp\shakespeare\collocates"
tells WordSmith to compute the collocates of each word in the shakespeare.lst word-list, and save results as plain text files, one per word, in the c:\temp\shakespeare\collocates folder.
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.

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

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

5.30.1 search for word or part of word

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

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

To search again, press OK again.

**Whole word – or bung in an asterisk**

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

See also: Searching by Typing, Search & Replace.

5.30.2 search by typing

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

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

5.30.3 search & replace

Some lists, such as lists of filenames, 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.

Search & Replace for filenames

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

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

![Search and Replace -- filenames](image)

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

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

Search & Replace for other data

In this case the search & replace isn't of filenames but in the case below in Viewer and Text Aligner,
of the actual text. Like a search 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.

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

### 5.31 filenames tab

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

See also: finding source files.
5.32 settings

5.32.1 save defaults

Settings can be altered by choosing Colour Settings in the WordSmith Tools Controller. 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** set WordSmith so it "knows" which folders you usually use
- **folders** character set, treatment of hyphens & numbers, default file extension
- **languages** restore last file, printing
- **general** tags to ignore, tag file, tag file autoloading, custom tagsets
- **tags** tags lists for Concord, KeyWords and Wordlist matching
- **stop lists** for Concord, KeyWords and Wordlist matching
- **Concord** number of entries, sort system, collocation horizons
- **KeyWords** procedure, max. p value, database & associate minimum frequencies, reference corpus filename
- **WordList** word length & frequencies, type/token #, cluster settings
- **Index** making a word-list index
- **Advanced** advanced settings
- **WSConcgram** for the concgram utility

**permanent settings and wordsmith6.ini file**

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

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

You can restore the defaults.

**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 $\text{Documents}\backslash wsmith6\backslash sayings.txt$, which holds sayings that appear in the main Controller window, if you don't like the sayings or want to add some more.

**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 $\text{wordsmith6.ini}$, usually off the root folder of $\text{My Documents}$.

The first time you use WordSmith, you will be prompted to choose appropriate Folders, Text, Characteristics, Tag 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 $\text{wordsmith6.ini}$, where it may be both read and altered. If WordSmith Tools finds a copy of $\text{wordsmith6.ini}$ in that folder it will be able to use your personal preferences.

### 5.32.2 restoring settings

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

**The point of it...**
You may have changed settings and cannot recall how to undo them....
Factory Defaults
Restores your `wordsmith6.ini` file to factory condition. Re-starts WordSmith with all relevant boxes filled in accordingly.

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

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

Colours
Re-sets colours to the factory defaults.

5.33 source text(s)

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

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

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

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

Menu options (right-click to see these)

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

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

- **Grey markup, Clear markup, Restore markup**
  Grey markup lets you grey out all `< >` sections.
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:

```
<table>
<thead>
<tr>
<th>ACTUALLY</th>
<th>1,846</th>
<th>YOU</th>
<th>21,331</th>
</tr>
</thead>
<tbody>
<tr>
<td>THINK</td>
<td>4,454</td>
<td>YEAH</td>
<td>4,185</td>
</tr>
<tr>
<td>WE</td>
<td>13,099</td>
<td>I</td>
<td>27,431</td>
</tr>
<tr>
<td>THAT</td>
<td>37,605</td>
<td>THAT'S</td>
<td>3,661</td>
</tr>
<tr>
<td>COUNCILLOR</td>
<td>608</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

...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
5.34 stop lists

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

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

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

Example

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

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

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.

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

**continuous**

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

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

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

**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, Match List, Lemmatisation.

5.35 suspend processing

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

- **carry on**
  ... as if you had not interrupted anything

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

**stop as soon as possible**

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

**panic stop**

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

Press *Suspend* again to effect your choice.

### 5.36 text and languages

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

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

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

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

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

Examples:
- ‘ (only apostrophes allowed in the middle of a word)
- % (both apostrophes and percent symbols allowed in the middle of a word)
- _ (both apostrophes and underscore characters allowed in the middle of a word)

You can include up to 10.

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

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

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

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

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

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

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

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

### 5.37 text dates and time-lines

**The point of it ...**

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

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

This screenshot shows a time-line based on concordancing *energy/emissions/carbon* in about 30 million words of UK newspaper text dealing with climate change, 2000-2010.
The first line shows overall data where all results on three search-terms are merged. Concordance hits are represented as a graph with green lines and little red blobs for each time period. The grey rectangles and the grey graph line both represent the same background information, namely the amount of word-data searched. The difference is merely that the grey line is twice as high as the rectangles below it. The number of hits in each year is mostly roughly proportional to the amount of text being examined, though in 2006 and 2009 for the term emissions it seems that the hit rate was slightly higher. In the first half of the decade carbon was rather under-mentioned in proportion to the amount of climate-change data being studied.

See also choosing text files: setting file dates

5.38 window management

The main WordSmith Tools Controller will be at the top left corner of your screen, half the screen width and half the screen height in size. Other Tools will appear in the middle. Each Tool main window will come just below any previous ones. Make use of the Taskbar (or Alt-tab, which helps you to switch easily from one window to the next).

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

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

tile and cascade
You can Tile or Cascade the Tools from the main WordSmith Tools program.

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

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

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 cash 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 a novel, *Bleak House* (Charles Dickens). The highlighted word *Guppy* is the name of one of the characters and details of this word are shown to the right.

**What you can see and do**

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

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

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

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

### 5.40 zap unwanted lines

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

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

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

See also : *reduce data to N entries*
Tags and Markup

Section VI
6 Tags and Markup

6.1 overview

What is markup for?

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

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

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

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 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, Making a Tag File, Showing Nearest Tags in Concord, Concord Sound and Video, Tag Concordancing, Types of Tag, Viewing the Tags, Using Tags as Text Selectors, Tags in WordList, XML text.
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. <>>&lt; 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 which lists all those you want to keep. These will then appear in your word lists and concordances.

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

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

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

translate entity references into other characters
If you use XML, SGML or HTML tagged text, you may want to translate symbols. For example, SGML, XML, HTML use &mdash; instead of a long dash. To do this, first prepare a tag file 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
If you declare the appropriate filename in your defaults (wordsmith6.ini) and include autoload tagfile=YES (or autoload tags to exclude file=YES, or autoload tags to translate file=YES), the markup-file will be automatically loaded as WordSmith starts up.

See also: Guide to handling the BNC, Overview of Tags, Making a Tag File, Showing Nearest Tags in Concord, Tag Concordancing, Types of Tag, Viewing the Tags, Using Tags as Text Selectors, Tags in WordList
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\wsmith6 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.

See also: Tags as text selectors

### 6.4 tags as selectors

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

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

**Markup to ignore**

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

![Advanced Settings](image)

**Markup to INclude or EXclude**
Entity file

Text Files and Mark-up

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

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

Document Header

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

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

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

The order in which these choices are handled

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

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

See also: Overview of Tags, Making a Tag File, Tag Handling, Tag Concordancing, Showing Nearest Tags in Concord, Viewing the Tags, Types of Tag, Guide to handling the BNC, XML text

## 6.5 only if containing...

### The point of it

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

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

Knowing that the BNC uses stext> for spoken texts, sex=m for males, age=5 for speakers aged 60 or more, you can get WordSmith to filter your text selection. It will search through the whole of every text file (not just the tags or header sections, in fact the first 2 megabytes 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 and can save the list for later use as favourites.

### 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.

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

6.6 part of file: selecting within texts

The point of it

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

Cut out or Keep?

Press the Cut or Keep tab to choose to cut out certain sections, and/or only to use certain sections.
Sections to Cut
Note: if you only want to remove a document header such as `</header>`, it is easier to do that in the general tag settings, section Document Header.

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

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 text you'd get all the major headings in your texts, however many, but nothing else.

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

**Ignore text files not containing choices**

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

**OK**

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

![Text Files & Mark-up](image)

See also: [Tags as Selectors](#), [Only if containing <x>](#), [Guide to handling the BNC](#).

6.7 **making a tag file**

**Tag Syntax**

Each tag is case sensitive.

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

You can use

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

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

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

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

**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>`
- `<SCENE>*/<SCENE>`
won't work, because both parts of the tag must be on the same line.
- `<SCENE>*/<SCENE>`
won't work, because the slash must be `/`.

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

**Mark-up to INclude**

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

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

```
<noun> /colour="yellow on red"
```

Available colours:

- 'Black', 'White', 'Cream',
- 'Red', 'Maroon',
- 'Yellow',
- 'Navy', 'Blue', 'Light Blue', 'Sky Blue',
- 'Green', 'Olive', 'Dollar Green', 'Grey-Green', 'Lime',
- 'Purple', 'Light Purple',
- 'Grey', 'Silver', 'Light Grey', 'Dark Grey', 'Medium Grey'.

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

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

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

and a text occurrence like

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

or

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

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

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

```
<w NN*> /description "noun" /colour="Cream on Purple"
```

**Section tag**

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

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

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

Here is an example of what you see after selecting a tag file and pressing "Load". The first tag is a "play media" tag, as is shown by the icon. You can see the cream on purple colour for nouns too. The tag file (BNC World.tag) is included in your installation.
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\wsmith6\sgmltrns.tag) which is included with your installation has been loaded. You could make a new one by reading it in, altering it, and saving it under a new name.

See also: Overview of Tags, Handling Tags, Showing Nearest Tags in Concord, Tag Concordancing, Types of Tag, Viewing the Tags, Using Tags as Text Selectors, Guide to handling the BNC.

### 6.8 tag-types

You will need to specify how each tag type starts and ends, and you should be consistent in usage. Restrict yourself to symbols which otherwise do not appear in your texts.

**eight special markers**

Eight kinds of marker may be marked as significant for word lists: those which represent starts and ends of headings, sections, sentences and paragraphs. Type these in the appropriate spaces when selecting Text Characteristics.
tags within 2 separators

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

<Noun>, <Verb>, <Pronoun>.

entity references

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

Specify these two types of markup by choosing Settings/Tag Lists, or Settings/Text Characteristics/Tags. You will then see a dialogue box offering Text to Ignore and a Browse button.

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

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

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

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

multimedia markers

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

See also: Overview of Tags, Handling Tags, Making a Tag File, Showing Nearest Tags in Concord, Tag Concordancing, Viewing the Tags, Using Tags as Text Selectors, Concord Sound and Video, Guide to handling the BNC.

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

He_PRONOUN entered_VERB the_DET room_NOUN.

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

<PRONOUN>He <VERB>entered <DET>the <NOUN>room.)

6.9 start and end of text segments

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

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

**Paragraphs**
For example, `<p *> or `<p>` might represent the beginning of a paragraph and `</p>` the end.

**Headings**
For example, `<head>` might represent the beginning and `</head>` the end. Note that the British National Corpus marks sentences within headings. Eg.

```xml
<head>
<s n="2"><w>NN1</w>Introduction</head>
```

in text HXL. It seems odd for the one word Introduction to count as a sentence, so WordSmith does not use sentence-tags within headings.

**Sections**
For example, `<section *> might represent the beginning and `</section>` the end.

Each of these is counted preferably when its closing tag such as `</s>`, `</p>` etc. is encountered. If there are no closing `</p>` tags in the entire text then paragraphs will be counted each time the opening paragraph tag is found.

See also: Overview of Tags, Handling Tags, Showing Nearest Tags in Concord, Tag Concordancing, Types of Tag, Viewing the Tags, Using Tags as Text Selectors, Guide to handling the BNC.

### 6.10 multimedia tags

In this screenshot you see an example of how to define your multimedia tags. This is accessed from Main Settings | Advanced | Tags | Media Tags.
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/1/32401-1.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, `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**
Duration to play and where to start playing are measured in seconds.
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>
```
If so, you'd specify duration to play as `play=` and where in the media file as `start=`
You can specify a default duration as in the screenshot: 20 seconds. As much as this may be needed especially if the sound tags are not spaced closely together in the text file.
If no start or duration indication is given, the whole sound or video file will be played.
If there are no duration and start position markers, the first number will be interpreted as start position and the second as duration, so a tag like this: `<sound$$C:\mysounds\talk.wav 15 5>` in your text file means "play `C:\mysounds\talk.wav` starting 15 seconds from the beginning and play for 5 seconds". If there's only one number as in `<sound$$C:\mysounds\talk.wav 15>`, that means "play `C:\mysounds\talk.wav` starting 15 seconds from the beginning and play for the default number of seconds".

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

See also: Sound and Video in Concord, Overview of Tags, Making a Tag File, Tag Handling, Tag Concordancing, Showing Nearest Tags in Concord, Viewing the Tags, Types of Tag, Guide to handling the BNC

### 6.11 modify source texts

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

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

**User-defined categories**
For example, suppose you have marked your concordance lines' `Set column` like this:
where the first line with miracle pre-modifies the noun cure and is marked adjective 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:
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: user-defined categories

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

```xml
<s n="43">I mean, </s>
<c><w c5="PUN" hw="i" pos="PRON">I</w><w c5="VVB" hw="mean" pos="VERB">mean</w></c>
<c><w c5="AVQ" hw="where" pos="ADV">where</w></c>
```
do eating disorders come from?

WordSmith’s handling of XML

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

Searching using Attributes

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

<w c5="NN2" * * *>

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

You would get results like this:

Hide the mark-up

If you prefer not to see all that the mark-up in grey, choose to hide the undefined mark-up
There is a button in the main tool which can show or hide mark-up, too.

**Asterisks in your search-word**

In the example above, we search on

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

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.

A search on

```xml
<w c5="NN?" hw="player" *>*
```

where we are allowing NN1 or NN2 and requiring the `hw` to be `player`, gets results like this:
Another example

Searching Italian .XML containing text like this:

and wishing to find all cases of the ARTPRE part of speech, with the search-word specified like this

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
7 Concord

7.1 purpose

Concord is a program which makes a concordance using plain text or web text files. To use it you will specify a search word or phrase, which Concord will seek in all the text files you have chosen. It will then present a concordance display, and give you access to information about collocates of the search word, dispersion plots showing where the search word came in each file, cluster analyses showing repeated clusters of words (phrases) etc.

The point of it...

The point of a concordance is to be able to see lots of examples of a word or phrase, in their contexts. You get a much better idea of the use of a word by seeing lots of examples of it, and it's by seeing or hearing new words in context lots of times that you come to grasp the meaning of most of the words in your native language. It's by seeing the contexts that you get a better idea about how to use the new word yourself. A dictionary can tell you the meanings but it's not much good at showing you how to use the word.

Language students can use a concordancer to find out how to use a word or phrase, or to find out which other words belong with a word they want to use. For example, it's through using a concordancer that you could find out that in academic writing, a paper can describe, claim, or show, though it doesn't believe or want (*this paper wants to prove that ...).

Language teachers can use the concordancer to find similar patterns so as to help their students. They can also use Concord to help produce vocabulary exercises, by choosing two or three search-words, blanking them out, then printing.

Researchers can use a concordancer, for example when searching through a database of hospital accident records, to see whether fracture is associated with fall, grease, ladder. Or to examine historical documents to find all the references to land ownership.

Online step-by-step guide showing how

7.2 index

Explanations

What to do if it doesn't do what I want ... [44]  
What is Concord and what's it for? [15]  
Collocation [17]  
Collocation Display [17]  
Plots [15]
Clusters
Patterns
Settings
Choosing texts
Collocate horizons
Collocate settings
Concordance settings
Context word
Main Controller Concordance Settings
Nearest Tag
Search word or phrase
Tag Concordancing
Tagged Texts
Text settings

Procedures
What you can See and Do
Altering the View
Blanking Out a Concordance
Re-sorting a Concordance
Removing Duplicate lines
Re-sorting Collocates
User-defined categories
Editing Concordances
Merging Concordances
Sound and Video in Concord

see also : WordSmith Main Index

7.3 what is a concordance?

A concordancer searches through a text or a group of texts and then shows the concordance as output. This can be saved, printed, etc.

7.4 search-word or phrase

7.4.1 search word syntax

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

Basic Examples

<table>
<thead>
<tr>
<th>search word</th>
<th>finds</th>
</tr>
</thead>
<tbody>
<tr>
<td>book</td>
<td>Book or book or BoOk</td>
</tr>
</tbody>
</table>
### 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: \( \text{CHR}(\text{xxx}) \) where \text{xxx} is the decimal number of the code. Examples: \( \text{CHR}(13) \) is a carriage-return, \( \text{CHR}(10) \) is a line-feed, \( \text{CHR}(9) \) is a tab. To represent \textit{<Enter>} which comes at the end of paragraphs and sometimes at the end of each line, you'd type \( \text{CHR}(13) \) \( \text{CHR}(10) \) which is carriage-return followed immediately by line-feed.

\( \text{CHR}(34) \) refers to double inverted commas.

\( \text{CHR}(46) \) is a full stop. There is a list of codes at \url{http://www.asciitable.com/}

You can also use hex format for numbers, e.g. \#x9 for tab, \#x22 for double inverted commas.

**Tags**

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

Examples:

<table>
<thead>
<tr>
<th>symbol</th>
<th>meaning</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;w NN1&gt;*</td>
<td>single common noun (BNC)</td>
<td>book, chair, elephant</td>
</tr>
<tr>
<td>&lt;w NN1?&gt;*</td>
<td>singular or plural common noun</td>
<td>book, chairs</td>
</tr>
<tr>
<td>&lt;w NN1&gt; t*</td>
<td>any single noun beginning with \text{T} or \text{t}</td>
<td>table, teacher</td>
</tr>
<tr>
<td>&lt;w NN1&gt;* &lt;w NN1&gt;*</td>
<td>two single common nouns in sequence</td>
<td>campaign manager</td>
</tr>
</tbody>
</table>

for XML formats see \url{XML text handling}

See also: Tag Concordancing, Context Word, Modify source texts, Ignore punctuation, Wildcards

### 7.4.2 file-based search-words

**The point of it...**

To save time typing in complex searches.

You may want to do a standard search repeatedly on different sub-corpora.

Or as Concord allows an almost unlimited number of entries, you may wish to do a concordance involving many search-words or phrases.

The space for typing in multiple search-words is limited to 80 characters (including / etc.). If your preferred search-words will exceed this limit or you wish to use a standardised search, you can prepare a file containing all the search-words.

**How to do it...**
A sample (Documents\wsmith6\concordance_search_words.txt) is included with the distribution files.

Use a Windows editor (e.g. Notepad) to prepare your own. Each one must be on a separate line of your file. No comment lines can be included, though blank lines may be inserted for readability.

If you want to require a context for a given word, put context:= as in this example:

book context:=hotel

(which seeks book and only shows results if hotel comes in the context horizons).

Then, instead of typing in each word or phrase in the Search Word dialogue box, just browse for the file.

Then press Load to read the entries (or Clear if you change your mind).

Lemmas and file-based concordancing

Note that where Concord has been called up from WordList, and the highlighted word in the word list is the head entry with lemmas, a temporary file will be created, listing the whole set of lemmas, and Concord will use this file-based search-word procedure to compute the concordance.

The temporary file will be stored in your Documents\wsmith6 folder unless you’re running on a network in which case it’ll be in Windows’ temporary folder, e.g. \windows\temp. It’s up to you to delete the temporary file.

Automated file-based concordances

If you want Concord to process a whole lot of different search-words, saving each result as it goes along so you can get a lot of work done with WordSmith unattended, choose SW Batch under
7.4.3 search-word and other settings

**Search Word or Phrase and/or Tag**

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

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

```plaintext
this
tat
the other
==Major==
```

Press the file button to locate your text file, the press the *Load* button. This will then change its name to something like *Clear 4*, where 4 means as in the example above that there are 4 different search-words to be concordanced. See "Batch" below for details on saving each one under a separate filename, otherwise all the searches will be combined into the same concordance.

---

**Advanced searches**
lemma list search
This option requires you to have chosen and loaded a lemma file. 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. Or if the search word is `book*` and the exclude if box has `hotel`, you'll get `book, books, booked, booking, bookable`, as long as `hotel` is not found within your context search horizons. Or if the search word is `*ish` and the exclusion specifies `fish`, you'll get `yellowish, greenish`, etc. but not `fish`.

You may type tag mark-up in here too, e.g. search for `book` with a context word `<ADJ>*` in position up to L3 will find book with a preceding adjective if your text has that sort of mark-up and if you've defined a tag file including `<ADJ>`.

In the screenshot above you see that "stop at sentence break" has been selected, meaning that a collocation search will only go left or right of the search-word up to a sentence-end. This is further explained here.
**Batch**

Suppose you're concordancing book* in 20 text files: you might want One concordance based on all 20 files (the default), or instead 20 separate concordances in a zipped batch which can be viewed separately (Text Batch). If you have multiple search-words in a file-based search as explained above, you may want each result saved separately (SW Batch).

Other settings affecting a concordance are available too:

see [WordSmith Controller Concordance Settings](#); [Typing characters](#)
[Accented characters](#); [Choosing Language](#); [Context Word(s) & Context Search Horizons](#)

### 7.5 advice

You have a listing showing all the concordance lines in a window. You can scroll up and down and left or right with the mouse or with the cursor keys.

#### Sort the lines

If you have a lot of lines you should certainly sort them. A concordance is read vertically, not horizontally. You are looking for repeated patterns, such as the presence of lots of the same sorts of words to the right or left of your search-word. Click the bar at the top to start a sort.
The Columns

These show the details for each entry: the entry number, the concordance line, set, tag, word-position (e.g. 1st word in the text is 1), paragraph and sentence position, source text file name, and how far into the file it comes (as a percentage). See below for an explanation of the purple blobs. The statistics are computed automatically based on the language settings.

- **Set**
  This is where you can classify the entries yourself, using any letter, into user-defined categories. Supposing you want to sort out verb uses from noun uses, you can press V or N. To type more (eg. "Noun"), double-click the entry in the set column and type what you want. If you have more than one search-word, you will find the Set column filled with the search-word for each entry. To clear the current entry, you can type the number 0. To clear the whole Set column, choose Edit | Clear Set column.

- **Tag**
  This column shows the tag context.

More context?

- **Stretching the display to see more**
  You can pull the concordance display to widen its column. Just place the mouse cursor on the bar between one column and another; when the cursor changes shape you can pull the whole column.

- **Stretch one line to see more context**
  The same applies to each individual row: place the mouse cursor between one row and another in the grey numbered area, and drag.
  Or press (F8) to "grow" all the rows, or (Ctrl+F8) to shrink them. Or press the numeric key-pad 8 to grow the current line as shown below. (Use numeric key-pad 2 to shrink it.)
Viewing the original text-file

(if it is still on the disk where it was when the concordance was originally created)

Double-click the concordance column, and the source text window will load the file and highlight the search word.

Or double-click the filename column, it will open in Notepad for editing.

Other things you may wonder about

Weird purple marks

In the screenshot you will see purple marks where any column is not wide enough to show all the data. The reason is that numbers are often not fully visible and you might otherwise get the wrong impression. For example in the concordance below, the Word # column shows 4,569 but the true number might be 14,569. Pull the column wider and the purple lines disappear.

Status bar

The status bar panels show:

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

See also:

Re-sorting your concordance lines
Follow-up searches
User-defined categories
Altering the View
Blanking out 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* (words in the neighbourhood of the search-word)

*Plot* (plots where the search-word came in the texts)

*Clusters* (groups of words in your concordance)

*Text segments in Concord* (paragraphs where the search-word came in the texts)

*Editing the concordance* (to edit a concordance)

*Time-lines* (to display a time-line)

*Zapping entries* (to remove entries from a concordance)

*Saving and printing* (to save and print your concordance)

*Window Management* (to manage your concordance windows)

7.6 **blanking**

In a concordance, to blank out the search-words with asterisks, just press the spacebar (or choose *View* / *Blanked out*). Press it again to restore them.

**The point of it...**

A blanked-out concordance is useful when you want to create an exercise. This one has *give* and *put* mingled:

... could not ********** me the time ...
... Rosemary, ********** me another ...
... would not ********** much for that ...
... could not ********** up with him ...
... so you'll ********** him a present ...
... will soon ********** up smoking ...
... he should ********** it over here ...

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

See also: *Other main Controller settings for Concord* [177]

7.7 **Category or Set**

7.7.1 **set column categories**

**The point of it...**

You may want to classify entries in your own way, e.g. separating adjectival uses from nominal ones, or sorting according to different meanings.
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 to include this annotation work you've done.

**How to do it**

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

You can sort 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 them, then choose Set column | Edit.
then type in a suitable value.

**Colours**

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

![Colours dialog box](image)

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.)

See also: Colour categories, modify your source texts, edit v. type-in mode.

7.7.2 colour categories

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

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

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

7.8 clusters

The point of it...

These word clusters help you to see patterns of repeated phraseology in your concordance, especially if you have a concordance with several thousand lines. Naturally, they will usually contain the search-word itself, since they are based on concordance lines. Another feature in Concord which helps you see patterns is Patterns.

How it does it...

Clusters are computed automatically if this is not disabled in the main Controller 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, the idea of “stop at sentence breaks” (there are other alternatives) is that a cluster which spans across two sentences is not likely to make sense.

### Re-computing clusters

The default clusters computed may not suit, (and you may want to recompute after deleting some lines), so you can also choose Compute / Clusters in the Concord menu, so as to choose how many words a cluster should have (cluster size 2 to 4 words recommended), and alter the other settings.
When you press OK, clusters will be computed. In this case we have asked for 3- 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:
"Related" clusters are those which overlap to some extent with others, so that the cause of overlaps with devoted to the cause of, etc. The procedure seeks out cases where the whole of a cluster is found within another cluster.

See also: general information on clusters, WordList Clusters, Word Clouds.
7.9 Collocation

7.9.1 what is collocation?

What's a "collocate"?
Collocates are the words which occur in the neighbourhood of your search word. Collocates of letter might include post, stamp, envelope, etc. However, very common words like the will also collocate with letter.

and "colligation"?
Linkages between neighbouring words which involve grammatical items are often referred to as colligation. That rely is typically followed by a preposition in English is a colligational fact.

The point of it...
The point of all this is to work out characteristic lexical patterns by finding out which "friends" words typically hang out with. It can be hard to see overall trends in your concordance lines, especially if there are lots of them. By examining collocations in this way you can see common lexical and grammatical patterns of co-occurrence.

Options
You may compute a concordance with or without collocates (minimal processing): without is slightly quicker and will take up less room on your hard disk. The default is to compute with collocates.
The number of collocates stored will depend on the collocation horizons. You can re-compute collocates after editing your concordance.
If you want to filter your collocate list, use a match list or stop list.
Re-sort a collocate list in a variety of ways.
You can see the strength of relationship between the word and the search-word which the concordance was based on.

Collocates can be viewed after the concordance has been computed.

Technical Note

The literature on collocation has never distinguished very satisfactorily between collocates which we think of as "associated" with a word (letter - stamp) on the one hand, and on the other, the words which do actually co-occur with the word (letter - my, this, a, etc.). We could call the first type "coherence collocates" and the second "neighbourhood collocates" or "horizon collocates". It has been suggested that to detect coherence collocates is very tricky, as once we start looking beyond a horizon of about 4 or 5 words on either side, we get so many words that there is more noise than signal in the system.

KeyWords allows you to study Associates, which are a pointer to "coherence collocates". Concord will supply "neighbourhood collocates". WordList and Concord allow you also to study relationships between words.

See also: collocation display, collocation settings, collocation relationship, relationships between words display.
7.9.2  collocate horizons

The collocate horizons represent the number of collocates Concord will find to the left and right of your search word, and the distance used by KeyWords in searching out plot-links. The default 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 too.

To set collocation horizons and other Concord settings, choose Concord Main Controller Settings or in the main WordSmith Controller choose Concord Settings.

See also: Collocate Settings

7.9.3  collocation relationship

The point of it...
The idea is to find out how strongly each collocate relates to the search-word near which it was found. MI (or other relevant statistic) is not computed by default for a collocate list.

How to compute it

In the Concord menu, choose Compute | Relationships:

Steps

1. Suppose you have made a concordance using all the files in Documents\wsmith6\text\shakespeare and have done a concordance on love. You see collocates such as Romeo, hate, the, Juliet, Nurse etc. All these show a "Relation" score of "??" because they haven't yet been computed.

2. If you haven't done so yet, use WordList to make a word list of the same text files (or if you prefer, use some other reference corpus). Make sure the reference corpus file is what you prefer.

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

You can choose a different statistic in the main Controller Concord settings.

Note: the procedure goes through your collocates and tries to find each in the word-list. If absent, you get a blank result. If one of your search-terms has a space in it such as Friar Lawrence, an ordinary single-word word list won't know its frequency and you will be asked to supply it. If you don't know, you should compute a concordance on that search-phrase over the same corpus first.
Full lemma processing, case sensitive
These are only relevant if your word list has any lemmatised entries, or it is a case-sensitive word list and you wish processing to respect case-sensitivity.

Relation statistic
Choose which type of relation you wish to compute. The default is Specific Mutual Information but in the screenshot Z score has been chosen.

Column for relation
The default is "Total". If you choose Total you're computing the relationship across the current collocation horizons set.
If you prefer to examine the relationship at only one position instead, you may:

See also: Collocation, Collocate display, Mutual Information

7.9.4 collocates display

Display
The collocation display initially shows the collocates in frequency order.
Beside each word and the search-word which the concordance was based on, you'll see the
strength of relationship between the two (or 0.000 if it hasn't yet been computed).
Then, the total number of times it co-occurred with the search word in your concordance, and a
total for Left and Right of the search-word. Then a detailed break-down, showing how many times it
cropped up 5 words to the left, 4 words to the left, and so on up to 5 words to the right. The centre
position (where the search word came) is shown with an asterisk.

The number of words to left and right depends on the collocation horizons. The numbers are:
- the total number of times the word was found in the neighbourhood of the search word
- the total number of times it came to the left of the search-word
- the total number of times it came to the right of the search-word
- a set of individual frequencies to the left of the search word (5L, i.e. 5 words to the left, 4L .. 1L)
- a Centre column, representing the search-word
- a set of individual frequencies to the right of the search word (1R, 2R, etc.)

The number of columns will depend on the collocation word horizons. With 5,5 you'll get five
columns to the left and 5 to the right of the search word. So you can see exactly how many times
each word was found in the general neighbourhood of the search word and how many times it was
found exactly 1 word to the left or 4 words to the right, for example.
The most frequent will be signalled in most frequent collocate colour (default=red). In the
screenshot below, differences comes 44 times in total but 39 of these are in position L1.

![Collocation Results](attachment:collocation_results.png)

The screenshot above shows collocation results for a concordance of BETWEEN/AMONG sorted by
the Relation column, where items like differentiate, difference etc. are found to be
most strongly related to between. Further down the listing, some links concerning among
(growing, refugees) are to be seen.
The frequency display can be re-sorted and you can recalculate the collocates if you zap entries from the concordance or change the horizons.

You can also highlight any given collocate in your concordance display.

See also: Word Clouds, Collocation, Collocation Relationship, Collocates and Lemmas, Mutual Information.

### 7.9.5 collocates and lemmas

In the following case a lemma list 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.
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).

The red mark by BE (row 5) shows that this row gives collocation numbers covering all forms of BE such as WAS, WERE etc. Similarly, HAVE and A are lemmatised in this screenshot.

Thus, for your search-word and its variants you can see detailed frequencies, but its collocates, though they do get lemmatised, do not show you the variant forms or any specific frequencies.

7.9.6 collocate follow

The point of it

The idea (from Paul Raper) is to be able to follow up a collocate by requesting a new concordance based on it, in the same text files as selected for the collocate. This aids exploration of related words.

How to do it

Here is an example, where there is a collocate list relating to BEERS. Select a word of interest, such
as KEG,

and select **Follow collocate** in the Compute menu.

WordSmith starts up a new Concord window with a search on KEG.
The search is carried out on the most recently selected text files (selected using the file-choose window or by reading in a saved concordance).

### 7.9.7 collocate highlighting in concordance

**The point of it...**

The idea is to be able to see a selected collocate highlighted in the concordance. In this example, the texts were Shakespeare plays and search word was **love**. One of the collocates is **know**, occurring a total of 50 times, with the most frequent at position 4 words to the left of **love**.

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Total</th>
<th>Left</th>
<th>Right</th>
<th>L5</th>
<th>L4</th>
<th>L3</th>
<th>L2</th>
<th>L1</th>
<th>Centre</th>
<th>R1</th>
<th>R2</th>
<th>R3</th>
<th>R4</th>
<th>R5</th>
</tr>
</thead>
<tbody>
<tr>
<td>53</td>
<td>KNOW</td>
<td>50</td>
<td>9</td>
<td>14</td>
<td>2</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>54</td>
<td>WHICH</td>
<td>49</td>
<td>14</td>
<td>35</td>
<td>3</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>17</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>55</td>
<td>THEY</td>
<td>49</td>
<td>26</td>
<td>23</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>0</td>
<td>15</td>
<td>0</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>56</td>
<td>TRUE</td>
<td>49</td>
<td>36</td>
<td>14</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>18</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>57</td>
<td>WOULD</td>
<td>46</td>
<td>16</td>
<td>31</td>
<td>4</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>7</td>
<td>8</td>
<td>2</td>
</tr>
</tbody>
</table>

Double-clicking 14 in the L4 column to the right of **know**, we see this in the concordance:
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:
    - Total Left: those to the left (33 instances of *know* above)
    - Total Right: those to the right (17)
  - otherwise: those in that column only

### To get rid

Re-sort in a different way or choose the menu item `View | Refresh`.

### 7.9.8 collocate settings

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

Collocates are computed case-insensitively (so *my* in the concordance line will be treated like *My*). If you don't want certain collocates such as THE to be included, use a `stop-list`.

You can lemmatise (join related forms like *SPEAK* -> *SPEAKS, SPOKE, SPOKEN*) using a
lemma list file.

**Minimum Specifications**

The minimum length is 1, and minimum frequency is 1 (default is 10). You can specify here how frequently it must have appeared in the neighbourhood of the Search Word. Words which only come once or twice are less likely to be informative. So specifying 5 will only show a collocate which comes 5 or more times in the neighbouring context. Similarly, you can specify how long a collocate must be for it to be stored in memory, e.g. 3 letters or more would be 3.

**Horizons**

Here you specify how many words to left and right of the Search Word are to be included in the collocation search: the size of the "neighbourhood" referred to above. The maximum is 25 left and 25 right. Results will later show you these in separate columns so you can examine exactly how many times a given collocate cropped up say 3 words to the left of your Search Word. The most frequent will be signalled in the *most frequent collocate colour* (default=red).

**Breaks**

These are

- no limits
- stop at punctuation
- stop at sentence break
- stop at paragraph break
- stop at heading break
- stop at section break
- stop at end of text

which you will see in the bottom right corner of the screen visible in the Controller Concord Settings.

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):
7.9.9 re-sorting: collocates

The point of it...

is to home in, for example, on the ones in L1 or R1 position. To find sub-patterns of collocation, so as to more fully understand the company your search-word keeps.

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, 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 of your sorted column. In the screenshot below, a concordance on cash generated these R1 collocates (with most function words eliminated using a stoplist):

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>R1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>FLOW</td>
<td>303</td>
</tr>
<tr>
<td>2</td>
<td>FLOWS</td>
<td>166</td>
</tr>
<tr>
<td>3</td>
<td>BOOK</td>
<td>41</td>
</tr>
<tr>
<td>4</td>
<td>UNDERWRITTEN</td>
<td>37</td>
</tr>
<tr>
<td>5</td>
<td>ACCOUNTING</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>MARKET</td>
<td>31</td>
</tr>
<tr>
<td>7</td>
<td>LIMITS</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>RATIO</td>
<td>24</td>
</tr>
<tr>
<td>9</td>
<td>POSITION</td>
<td>22</td>
</tr>
<tr>
<td>10</td>
<td>PAYMENTS</td>
<td>20</td>
</tr>
<tr>
<td>11</td>
<td>RESERVES</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>INFLOWS</td>
<td>19</td>
</tr>
<tr>
<td>13</td>
<td>ALTERNATIVE</td>
<td>18</td>
</tr>
<tr>
<td>14</td>
<td>COWS</td>
<td>18</td>
</tr>
<tr>
<td>15</td>
<td>BALANCES</td>
<td>17</td>
</tr>
<tr>
<td>16</td>
<td>SURPLUSES</td>
<td>15</td>
</tr>
<tr>
<td>17</td>
<td>BASIS</td>
<td>14</td>
</tr>
<tr>
<td>18</td>
<td>PAYMENT</td>
<td>13</td>
</tr>
<tr>
<td>19</td>
<td>TRANSFER</td>
<td>13</td>
</tr>
<tr>
<td>20</td>
<td>DEPOSITS</td>
<td>12</td>
</tr>
<tr>
<td>21</td>
<td>EQUIVALENTS</td>
<td>12</td>
</tr>
<tr>
<td>22</td>
<td>OFFER</td>
<td>12</td>
</tr>
<tr>
<td>23</td>
<td>TRANSFERS</td>
<td>12</td>
</tr>
<tr>
<td>24</td>
<td>RESOURCES</td>
<td>11</td>
</tr>
<tr>
<td>25</td>
<td>INDEX</td>
<td>10</td>
</tr>
<tr>
<td>26</td>
<td>RECEIVED</td>
<td>10</td>
</tr>
</tbody>
</table>

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, Collocation Display, Collation Horizons, Word Clouds, Patterns.

### 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.

**What you see**

The plot shows:

- **File**: source text file-name
- **Words**: number of words in the source text
- **Hits**: number of occurrences of the search-word
per 1,000 how many occurrences per 1,000 words

Dispersion the plot dispersion value

Plot a plot showing where they cropped up, where the left edge of the plot represents 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.).

Here we see a plot of "O" and another of "AH" from the play Romeo and Juliet. They are on separate lines because there were 2 search-words. There are more "O" exclamations than "AH"s.

As the status bar says, you can get the word numbers for the plot by double-clicking the plot area:

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:

<table>
<thead>
<tr>
<th>File</th>
<th>Words</th>
<th>Hits</th>
<th>per 1,000</th>
<th>Dispersion</th>
<th>Plot</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Overall</td>
<td>562.64</td>
<td>182</td>
<td>0.33</td>
<td>0.528</td>
</tr>
<tr>
<td>2</td>
<td>a30.txt</td>
<td>6.632</td>
<td>1</td>
<td>0.15</td>
<td>-0.069</td>
</tr>
<tr>
<td>3</td>
<td>a50.txt</td>
<td>4.113</td>
<td>1</td>
<td>0.24</td>
<td>-0.069</td>
</tr>
<tr>
<td>4</td>
<td>a1.txt</td>
<td>8.643</td>
<td>2</td>
<td>0.23</td>
<td>0.300</td>
</tr>
<tr>
<td>5</td>
<td>a51.txt</td>
<td>3.815</td>
<td>3</td>
<td>0.31</td>
<td>0.250</td>
</tr>
<tr>
<td>6</td>
<td>a17.txt</td>
<td>8.747</td>
<td>1</td>
<td>0.11</td>
<td>-0.069</td>
</tr>
<tr>
<td>7</td>
<td>a28.txt</td>
<td>10.979</td>
<td>2</td>
<td>0.18</td>
<td>-0.300</td>
</tr>
<tr>
<td>8</td>
<td>a24.txt</td>
<td>6.065</td>
<td>1</td>
<td>0.17</td>
<td>-0.069</td>
</tr>
<tr>
<td>9</td>
<td>a2m.txt</td>
<td>9.507</td>
<td>2</td>
<td>0.21</td>
<td>0.300</td>
</tr>
<tr>
<td>10</td>
<td>a2u.txt</td>
<td>5.506</td>
<td>2</td>
<td>0.36</td>
<td>0.300</td>
</tr>
<tr>
<td>11</td>
<td>a35.txt</td>
<td>5.839</td>
<td>1</td>
<td>0.17</td>
<td>-0.069</td>
</tr>
<tr>
<td>12</td>
<td>a5c.txt</td>
<td>8.118</td>
<td>2</td>
<td>0.25</td>
<td>0.300</td>
</tr>
<tr>
<td>13</td>
<td>a3p.txt</td>
<td>7.022</td>
<td>3</td>
<td>0.35</td>
<td>0.475</td>
</tr>
<tr>
<td>14</td>
<td>a7.txt</td>
<td>8.047</td>
<td>1</td>
<td>0.25</td>
<td>-0.069</td>
</tr>
<tr>
<td>15</td>
<td>a3.txt</td>
<td>4.386</td>
<td>2</td>
<td>0.46</td>
<td>0.300</td>
</tr>
<tr>
<td>16</td>
<td>a5t.txt</td>
<td>5.137</td>
<td>1</td>
<td>0.19</td>
<td>-0.069</td>
</tr>
<tr>
<td>17</td>
<td>a5.txt</td>
<td>8.514</td>
<td>1</td>
<td>0.12</td>
<td>-0.069</td>
</tr>
<tr>
<td>18</td>
<td>a3b.txt</td>
<td>7.393</td>
<td>4</td>
<td>0.54</td>
<td>0.429</td>
</tr>
</tbody>
</table>

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:

![Concord settings](image)

and re-sort the File list:
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 and can get your spreadsheet to make graphs etc, as explained here.

Each plot window is dependent on the concordance from which it was derived. If you close the original concordance down, it will disappear. You can Print the plot. There's no Save option for the plot alone but you can of course save the concordance itself. You can Copy to the clipboard (Ctrl+C) and then put it into a word processor as a graphic, using Paste Special.
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.

Here is a set of preferences with lots of files and two search-words:
giving results like this:

![Image of WordSmith Tools interface]

See also: plot and ruler colours, plot dispersion value.

### 7.11 concordancing on tags

**The point of it...**

Suppose you're interested in identifying structures of a certain type (as opposed to a given word or phrase), for example sequences of **Noun+Noun+Noun**. You can type in the tags you want to concordance on (with or without any words).

**How to do it...**

In Concord's search-word box, type in the tags you are interested in. Or define your tags in a tag-file.

**Examples**

- `<w NN1>table` finds *table* as a singular noun (as opposed to as a verb)
- `<w NN1>* <w NN1>*` will find any sequence of two singular common nouns in the BNC Sampler.

Note that `<w NN1>table` finds *table* if your text is tagged with `<` and `>` symbols, or if you have specified `[` and `]` as tag symbols, it will find `[w NN1]table`.

There are some more examples under Search word or phrase.

It doesn't matter whether you are using a tag file or not, since WordSmith will identify your tags automatically. (But not by magic: of course you do need to use previously tagged text to use this function.)

In example 2, the asterisks are because in the BNC, the tags come immediately before the word they refer to: if you forgot the asterisk, Concord would assume you wanted a tag with a separator on either side.

**Are you concordancing on tags?**
If you are asked this and 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.

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 the mark-up. See the View menu for this.

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

7.11.1 nearest tag

Concord allows you to see the nearest tag, if you have specified a tag file, 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 with your sections and label them as sections, as in these examples:

<?xml version="1.0"?>
<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>
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 with each concordance line.

You can sort on the nearest tags. In the shot below, a concordance of such has been computed using BNC text. Some of the cases of such are tagged < PRP> (such as) and others are <w DT0>. The Tag column shows the nearest tag, and the whole list has been sorted using that column.

If you can't see any tags using this procedure, it is probably because the Tags to Ignore have the same format. For example, if Tags to Ignore has <">, any tags such as <title>, <quote>, etc. will be cut out of the concordance unless you specify them in a tag file. If so, specify the tag file and run the concordance again.

You can also display tags in colour, or even hide the tags -- yet still colour the tagged word. Here is a concordance of this in the BNC text with the tags in colour:
and here is a view showing the same data, with View | Hide Tags selected.
The tags themselves are no longer visible, and only 6 types of tag have been chosen to be viewed in colour.

See also: Guide to handling the BNC, Overview of Tags, Handling Tags, Making a Tag File, Tagged Texts, Types of Tag, Viewing the Tags, Using Tags as Text Selectors

7.12 context word

You may restrict a concordance search by specifying a context word which either must or may not be present within a certain number of words of your search word. For example, you might have book as your search word and hotel* as the context word. This will only find book if hotel or hotels is nearby.

Or you might have book as your search word and paper* as an exclusion criterion. This will only find book if paper or papers is not within your Context Search Horizons.

Context Search Horizons

The context horizons determine how far Concord must look to left and right of the search word when checking whether the search criteria have been met. The default 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`,

- `*` 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 `*`, `?`, `==`, `~`, `\`, `/` as a character in your search word, put it in double quotes, e.g. `"*"`
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:
If you have specified a context word, you can re-sort on it. Also, the context words will be in their own special colour.

Note: the search only takes place within the current concordance line with the number of characters defined as characters to save. That is, if for example you choose search horizons 25L and 25R, but only 1000 characters are saved in each line, there might not be 25 words on either side of the search-word to examine when seeking the context word or phrase if there was extensive mark-up as well.

7.13 editing concordances

The point of it...

You may well find you have got some entries which weren’t what you expected. Suppose you have
done a search for SHRIMP*/PRAWN* -- you may find a mention of Shrimpton in the listing. It's easy to clean up the listing by simply pressing Del on each unwanted line. (Do a sort on the search word first so as to get all the Shrimptons next to each other.) The line will turn a light grey colour.

Pressing Ins will restore it, if you make a mistake. To delete or restore ALL the lines from the current line to the bottom, press the grey - key or the grey + key by the numeric keypad. When you have finished marking unwanted lines, you can choose (Ctrl+Z or ⌘) to zap the deleted lines.

If you're a teacher you may want to blank out the search words: to do so, press the spacebar. Pressing the spacebar again will restore it, so don't worry!

7.14 patterns

When you have a collocation window open, one of the tab windows shows "Patterns". This will show the collocates (words adjacent to the search word), organised in terms of frequency within each column. That is, the top word in each column is the word most frequently found in that position. The second word is the second most frequent.

In R1 position (one word to the right of the search-word love) there seem to be both intimate (thee) and formal (you) pronouns associated with love in Shakespeare. And looking at L1 position it seems that speakers talk more of their love for another than of another's love for them.

The minimum frequency and length for one of the words to be shown at all, is the minimum frequency/length for collocates.

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 in your concordance display.
7.15 remove duplicates

The problem
Sometimes one finds that text files contain duplicate sections, either because the corpus has become corrupted through being copied numerous times onto different file-stores or because they were not edited effectively, e.g. a newspaper has several different editions in the same file. The result can sometimes be that you get a number of repeated concordance lines.

Solution
If you choose Edit | Remove Duplicates, Concord goes through your concordance lines and if it finds any two where the stored concordance lines are identical, regardless of the filename, date etc. it will mark one of these for deletion. That is, it checks all the characters to save to see whether the two lines are identical. If you set this to 150 or so it is highly unlikely that false duplicates will be identified, since every single character, comma, space etc. would have to match.

Check before you zap...
At the end it will sort all the lines so you can see which ones match each other before you decide finally to zap the ones you really don't want.

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

Main Sort
The listing can be sorted by three criteria at once. A Main Sort on Left 1 (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](#).

**sorting by set (user-defined categories)**

You can also sort by set, if you have chosen to classify the concordance lines according to your own scheme, using letters from A to Z or a to z or longer strings. The sort will put the classified lines first, in category order, followed by any unclassified lines. See [Nearest Tag](#) for details of sorting by tags. See [colour categories](#) 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 wordlist. The result will be to sort those lines which contain “easy” (highly frequent) words at the top of the list.

**All**

By default you sort all the lines; you may however type in for example 5-49 to sort those lines only.

**Ascending**

If this box is checked, sort order is from A to Z, otherwise it’s from Z to A.

See also: WordList sort, KeyWords sort, Choosing Language

### 7.16.1 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
7.17 saving and printing

You can save the concordance (and its collocates & other dependent results if these were stored when the concordance was generated) either as a Text File (e.g. for importing into a word processor) or as a file of results which you can subsequently Open (in the main menu at the top) to view again at a later date. When you leave Concord you'll be prompted to save if you haven't already done so.

Saving a concordance allows you to return later and review collocates, dispersion plots, clusters.

You can Print using the Windows printer attached to your system. You will get a chance to specify the number of pages to print. The font will approximate the one you can see on your screen. If you use a colour printer or one with various shades of grey, the screen colours will be copied to your printer. If it is a black-and-white printer, coloured items will come in italics if your printer can do italics.

Concord prints as much of your concordance plus associated details as your printing paper settings allow, the edges being shown in Print Preview.

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 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
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 to get a concordance into Word or another application.

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 a schwa? Does the second vowel sound like "i" or "e" or "u" or a schwa?

How to do it...
If you have defined tags which refer to multimedia files, and if there are any such tags in the "tag-context" of a given concordance line, you can hear or see the source multimedia. The tag will be parsed to identify the file needed, if necessary downloading it from a web address, and then played.

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. If you wish to use a transcript and sound file format which is incompatible with the syntax described here, please contact us.
In this screenshot we see a concordance where there is a tag inserted periodically in the text file. To play the media file, press Control/M or choose File / Play media file, or double-click the Tag column.

Video files can be played if the free VLC Media Player is installed (see http://www.vlcapp.com/). This screenshot shows a concordance line with, in the Nearest Tag column, the mark-up saying that the source text and the video file have the same file-name (except that the latter ends .AVI and the former .TXT). A double-click on the 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).

See also: Multi-media Tag syntax, Handling Tags, Making a Tag File, Showing Nearest Tags in Concord, Tag Concordancing, Types of Tag, Viewing the Tags, Using Tags as Text Selectors, Tags in WordList

7.19 summary statistics

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

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

and press Count;

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

**Breakdown**

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

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

The results window will now show something like this

where it is clear that the singular *consequence* came 20 times in 20 different files, the first being file A3A.TXT. Further down you will find the results for *consequences*:
which appeared 103 times in 74 files, and that in the first of these, A1E.TXT, it came twice.

**Cumulative column**
see the explanation for WordList

**Load button**
see the explanation for count data frequencies

### 7.20 text segments in Concord

A concordance line brings with it information about which segment of the text it was found in.

In the screenshot below, a concordance on *year* was carried out; the listing has been sorted by Heading Position -- in the top 2 lines, *year* is found as the 3rd word of a heading. The advantage of this is that it is possible to identify search-words occurring near sentence starts, near the beginning of sections, of headings, of paragraphs.
You can toggle the numbers between raw numbers and percentages.

See also: Start and end of text segments.
7.21 viewing options

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), but you can set this to different values if you like. For further details, see Sorting a Concordance.

Show collocate zero frequencies
This toggles whether 0 or a blank (the default) is shown if a collocate frequency is zero.
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.

- Show full filename and path = sometimes you need to see the whole path but usually the filename alone will suffice.
- Cut redundant spaces = remove any double spaces
- Show sentence only = show the context only up to its left and right sentence boundaries
- Show raw numbers = show the raw data instead of percentages e.g. for sentence position
- Hide search-word = blank it out e.g. to make a guess-the-word exercise
- Pad search-word with spaces = insert a space to left and right of the search-word so it stands out better
- Hide undefined tags = hide those not defined in your tag file
- Hide tag file tags = hide all tags including undefined ones
- Hide words = show only the tags

Some of the options are visible here:
for example the sub-set visible shows an opportunity to blank out the search-word, to pad it with a space left & right, to shift the search-word left or right.

See also: Controller, What you get with choices, showing nearest tags, blanking out the search-word, viewing more context, growing/shrinking concordance lines.

7.22 WordSmith controller: Concord: settings

These are found in the main Controller marked Concord. This is because some of the choices -- e.g. collocation horizons -- may affect other Tools.
When you have computed a concordance, the Concord button will have a red number (showing how many Concord windows are in use) and at the bottom of the screen you will see an icon. Click that to see the list of files and their features.

**WHAT YOU GET and WHAT YOU SEE**

There are 2 tabs for settings affecting *What you get* in the concordance and *What you see* in the display. There is a screenshot at Concord: viewing options 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.)

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
**auto remove duplicates:** removes any lines where the whole concordance entry matches another. (This can happen if you have a corpus where news stories get re-published in different editions by different newspapers.)

**Ignore punctuation between words:** this allows a search for **BY ITSELF** to succeed where the text contains **...went by, itself**

**Characters to save**
Here is where you set how many characters in a concordance line will be stored as text as the concordance is generated. The default and minimum is 1000. This number of characters will be saved when you save your results, so even if you subsequently delete the source text file you can still see some context. If you grow the lines more text will be read in (and stored) as needed. There are [examples here](#).
Save as text search-word or context-word marker: here you can also specify markers for your search-word and context-word.

Collocates

By default, Concord will compute collocates as well as the concordance, but you can set it not to if you like (Minimal processing). For further details, see Collocate Horizons or Collocation.

The minimum frequency and length refer to the collocates to be shown in your listing. With the settings above, only collocates which occur at least 5 times and contain at least 1 character will be shown as long as they don't cross sentence boundaries.

If separate search words is checked and you have multiple search-terms, then you get collocates distinguishing between the different search-terms. If you want them amalgamated, clear this check-box.

Collocates relation statistic

Choose between Specific Mutual Information, MI3, Z Score, Log Likelihood. See Mutual Information Display for examples of how these can differ.
The options are explained at Concord: viewing options.

**Columns**

The *Columns to show/hide* list offers all the standard columns: you may uncheck ones you normally do not wish to see. This will only affect newly computed KeyWords data: earlier data uses the column visibility, size, colours etc already saved. They can be altered using the Layout menu option at any time.

See also: Concord Saving and Printing, Concord Help Contents, Collocation Settings.
8 KeyWords

8.1 purpose

This is a program for identifying the "key" words in one or more texts. Key words are those whose frequency is unusually high in comparison with some norm. Click here for an example.

The point of it…

Key-words provide a useful way to characterise a text or a genre. Potential applications include: language teaching, forensic linguistics, stylistics, content analysis, text retrieval.

The program compares two pre-existing word-lists, which must have been created using the WordList tool. One of these is assumed to be a large word-list which will act as a reference file. The other is the word-list based on one text which you want to study.

The aim is to find out which words characterise the text you're most interested in, which is automatically assumed to be the smaller of the two texts chosen. The larger will provide background data for reference comparison.

Key-words and links between them can be plotted, made into a database, and grouped according to their associates.

Online step-by-step guide showing how

8.2 index

Explanations

What is the Keywords program and what's it for?

How Key Words are Calculated

2-Word list Analysis

Key words display

Key words plot

Key words plot display

Plot-Links

Batch Analyses

Database of Key Key-Words

Associates

Clumps

Limitations

Settings and Procedures

Calling up a Concordance
8.3 ordinary two word-list analysis

The usual kind of KeyWords analysis. It compares the one text file (or corpus) you're chiefly interested in, with a reference corpus based on a lot of text. In the screenshot below we are interested in the key words of deer hunter story and we're using BNC as the reference corpus to compare with.

Choose Word Lists
In the dialogue box you will choose 2 files. The text file in the box above and the reference corpus file in the box below.
8.4 associate definition

An “associate” of key-word X is another key-word (Y) which co-occurs with X in a number of texts. It may or may not co-occur in proximity to key-word X. (A collocate would have to occur within a given distance of it, whereas an associate is “associated” by being key in the same text.)

For example, in a key-word database of Guardian newspaper text, wine was found to be a key word in 25 out of 299 stories from the Saturday “tabloid” page, thus a key key word in this section. The top associates of wine were: wines, Tim, Atkin, dry, le, bottle, de, fruit, region, chardonnay, red, producers, beaujolais.

It is strikingly close to the early notion of “collocate”.

Association operates in various ways. It can be strong or weak, and it can be one-way or two-way.

For example, the association between to and fro is one-way (to is nearly always found near fro but it is rare to find fro near to).

See also: Definition of Key Word, Associates, Definitions, Mutual Information
8.5 **associates**

“Associates” is the name given to key-words associated with a key key-word. 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:

- **Minimum texts**
  These screenshot settings would only process those key-key-words which appear in at least 3 text
files.

**Statistic**

Choose the mutual information statistic you prefer, apart from Z score which uses a span (here we're using the whole text).

**Minimum strength**

This will only show associates which reach at least the strength in the statistic set here, e.g. 3.000.

This screenshot shows the most frequent associates in the right-hand column of the main keywords data base window.

<table>
<thead>
<tr>
<th>N</th>
<th>KW</th>
<th>Texts</th>
<th>%</th>
<th>Freq</th>
<th>Ass.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>THE KING</td>
<td>94</td>
<td>10.00</td>
<td>933</td>
<td>94 lord lord cardinal john you advised the art, please your art, please my lord, lord of</td>
</tr>
<tr>
<td>2</td>
<td>MY LORD</td>
<td>86</td>
<td>16.00</td>
<td>621</td>
<td>76 lord of the king lord chamber lord cardinal, his grace, buckingham and my lord of</td>
</tr>
<tr>
<td>3</td>
<td>DUKE OF</td>
<td>5</td>
<td>13.00</td>
<td>140</td>
<td>47 his grace, he had good heed at court, a remnant, the gods, between you and the queen</td>
</tr>
<tr>
<td>4</td>
<td>FRANCE</td>
<td>5</td>
<td>13.00</td>
<td>86</td>
<td>48 king, Henry, in France, France, and a king and warnick the duke of York, of France</td>
</tr>
<tr>
<td>5</td>
<td>LORD OF</td>
<td>5</td>
<td>13.00</td>
<td>66</td>
<td>79 king, Henry, in France, France, and the king lord protector, lord cardinal, my lord of</td>
</tr>
</tbody>
</table>

To see the detailed associates, double-click your chosen term in the KW column or the Associates column:

**Associates of DUKE OF**

<table>
<thead>
<tr>
<th>N</th>
<th>Associate</th>
<th>Strength</th>
<th>Texts</th>
<th>%</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>OF FRANCE</td>
<td>5.755</td>
<td>3</td>
<td>60.00</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>OF YORK</td>
<td>5.755</td>
<td>3</td>
<td>60.00</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>THE DUKE OF</td>
<td>5.755</td>
<td>3</td>
<td>60.00</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>AND WARWICK</td>
<td>4.963</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>A KING</td>
<td>4.379</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>FRANCE AND</td>
<td>4.379</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>IN FRANCE</td>
<td>4.379</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>KING HENRY</td>
<td>4.379</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>OF ENGLAND</td>
<td>4.379</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>THE DAUPHIN</td>
<td>4.379</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>THE FRENCH</td>
<td>4.379</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>MY LORDS</td>
<td>3.963</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>MY LORD OF</td>
<td>3.642</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>THE DUKE</td>
<td>3.642</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>THE KING</td>
<td>3.156</td>
<td>2</td>
<td>40.00</td>
<td></td>
</tr>
</tbody>
</table>

See also: [definition of associate](#), [related clusters](#)
8.6 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 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.
8.7 clumps

"Clumps" is the name given to groups of key-words associated with a key key-word. The point of it (1)...

The idea here is to refine associates by grouping together words which are found as key in the same sub-sets of text files. The example used to explain associates will help. Suppose the word wine is a key key-word in a set of texts, such as the weekend sections of newspaper articles. Some of these articles discuss different wines and their flavours, others concern cooking and refer to using wine in stews or sauces, others discuss the prices of wine in a context of agriculture and diseases affecting vineyards. In this case, the associates of wine would be items like Chardonnay, Chile, sauce, fruit, infected, soil, etc. The associates procedure shows all such items unsorted.

The clumping procedure, on the other hand, attempts to sort them out according to these different uses. The reasoning is that the key words of each text file give a condensed picture of its "aboutness", and that "aboutnesses" of different texts can be grouped by matching the key word lists. Thus sets of key words can be clumped together according to the degree of overlap in the key word lexis of each text file.

Two stages

The initial clumping process does no grouping: you will simply see each set of key-words for each text file separately. To group clumps, you may simply join those you think belong together (by dragging), or regroup with help by pressing ?.
The listing shows clumps sorted in alphabetical order. You can re-sort by frequency (the number of times each key word in the clump appeared in all the files which comprise the clump).

See also: definition of associate, regrouping clumps

8.8 concordance

With a key word or a word list list on your screen, you can choose Compute and to call up a concordance of the currently selected word(s). The concordance will search for the same word in the original text file that your key word list came from.

The point of it…

is to see these same words in their original contexts.

8.9 creating a database

To build a key words database, you will need a set of key word lists. For a decent sized database, it is preferable to build it like this:

1. Make a batch of word lists.
2. Use this to make a batch of keyword lists. Set "faster minimal processing" on as in this shot, so as to not waste time computing plots etc.
3. Now, in KeyWords, choose New / KW Database.

This enables you to choose the whole set of key word files. Note that making a database means that only positive key words will be retained.

In the Controller KeyWords settings you can 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.
8.10 example of key words

You have a collection of assorted newspaper articles. You make a word list based on these articles, and see that the most frequent word is the. Among the rather infrequent words in the list come examples like hopping, modem, squatter, grateful, etc.

You then take from it a 1,000 word article and make a word list of that. Again, you notice that the most frequent word is the. So far, not much difference.

You then get KeyWords to analyse the two word lists. KeyWords reports that the most "key" words are: squatter, police, breakage, council, sued, Timson, resisted, community.

These "key" words are not the most frequent words (which are those like the) but the words which are most unusually frequent in the 1,000 word article. Key words usually give a reasonably good clue to what the text is about.

Here is an example from the play Othello.
8.11 key key-word definition

A "key key-word" is one which is "key" in more than one of a number of related texts. The more texts it is "key" in, the more "key key" it is. This will depend a lot on the topic homogeneity of the corpus being investigated. In a corpus of City news texts, items like bank, profit, companies are key key-words, while computer will not be, though computer might be a key word in a few City news stories about IBM or Microsoft share dealings.

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.

See also: How Key Words are Calculated, Definition of Key Word, Creating a Database, Definitions.

8.12 keyness

8.12.1 p value

(Default=0.000001)

The $p$ value is that used in standard chi-square and other statistical tests. This value ranges from 0 to 1. A value of .01 suggests a 1% danger of being wrong in claiming a relationship, .05 would give a 5% danger of error. In the social sciences a 5% risk is usually considered acceptable. In the case of key word analyses, where the notion of risk is less important than that of selectivity,
you may often wish to set a comparatively low $p$ value threshold such as 0.000001 (one in 1 million) (1E-6 in scientific notation) so as to obtain fewer key words. Or you can set a low "maximum wanted" number in the main Controller under KeyWords Settings.

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

See also: Definitions, choosing a reference corpus

8.12.2 key-ness definition

The term "key word", though it is in common use, is not defined in Linguistics. This program identifies key words on a mechanical basis by comparing patterns of frequency. (A human being, on the other hand, may choose a phrase or a superordinate as a key word.)

A word is said to be "key" if
a) it occurs in the text at least as many times as the user has specified as a Minimum Frequency
b) its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a $p$ value specified by the user.

positive and negative keyness

A word which is positively key occurs more often than would be expected by chance in comparison with the reference corpus.

A word which is negatively key occurs less often than would be expected by chance in comparison with the reference corpus.

typical key words

KeyWords will usually throw up 3 kinds of words as "key". First, there will be proper nouns. Proper nouns are often key in texts, though a text about racing could wrongly identify as key, names of horses which are quite incidental to the story. This can be avoided by specifying a higher Minimum Frequency.

Second, there are key words that human beings would recognise. The program is quite good at finding these, and they give a good indication of the text’s "aboutness". (All the same, the program does not group synonyms, and a word which only occurs once in a text may sometimes be "key" for a human being. And KeyWords will not identify key phrases unless you are comparing word-lists based on word clusters)

Third, there are high-frequency words like because or shall or already. These would not usually be identified by the reader as key. They may be key indicators more of style than of "aboutness". But the fact that KeyWords identifies such words should prompt you to go back to the text, perhaps with Concord (just choose Compute | Concordance), to investigate why such words have cropped up with unusual frequencies.

See also: How Key Words are Calculated, Definition of Key Key-Word, Definitions, KeyWords Settings
8.12.3 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.

8.13 KeyWords database

(default file extension .KDB)

The point of it...
The point of this database is that it will allow you to study the key-words which recur often over a number of files.
For example, if you have 500 business reports, each one will have its own key words. These will probably be of two main kinds. There will be key-words which are key in one text but are not
generally key (names of the firms and words relating to what they individually produce); and other, more general words (like consultant, profit, employee) which are typical of business documentation generally. Or you may find that I, you, should etc. come to the top if your text files are ones which are much more interactive than the reference corpus texts.

By making up a database, you can sort these out. The ones at the top of the list, when you view them, may be those which are most typical of the genre in some way. We might call the ones at the top "key-key words" and the list is at first ordered in terms of "key key-ness", but those at the bottom will only be key in a few text files. You can of course toggle it into alphabetical order and back again.

You can set a minimum number of files that each word must have been found to be key in, using KeyWords Settings | Database.

When viewing a database you will be able to investigate the associates of the key key-words. Under Statistics, you will also be able to see details of the key words files which comprise the database (file name and number of key words per file), together with overall statistics on the number of different types and the tokens (the total of all the key-words in the whole database including repeats).

See also: Creating a database, Definition of key key-word.

### 8.14 keywords 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.

### 8.15 KeyWords: advice

1. Don't call up a plot of the key words based on more than one text file. It doesn't make sense! Anyway the plot will only show the words in the first text file. If you want to see a plot of a certain word or phrase in various different files, use Concord dispersion.
2. There can be no guarantee that the "key" words are "key" in the sense which you may attach to "key". An "important" word might occur once only in a text. They are merely the words which are outstandingly frequent or infrequent in comparison with the reference corpus.

3. Compare apples with pears, or, better still, Coxes with Granny Smiths. So choose your reference corpus in some principled way. The computer is not intelligent and will try to do whatever comparisons you ask it to, so it's up to you to use human intelligence and avoid comparing apples with phone boxes!

**If it didn't work...**

For the procedure to work, a number of conditions must be right: the language 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 in ascending order before the comparison is made. (The program tries to ensure this, automatically.) Also, any prefixes or suffixes must match.

### 8.16 KeyWords: calculation

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

If the word occurs say, 5% of the time in the small word-list and 6% of the time in the reference corpus, it will not turn out to be "key", though it may well be the most frequent word. If the text concerns the anatomy of spiders, it may well turn out that the names of the researchers, and the items spider, leg, eight, etc. may be more frequent than they would otherwise be in your reference corpus (unless your reference corpus only concerns spiders!)

To compute the "key-ness" of an item, the program therefore computes

- its frequency in the small word-list
- the number of running words in the small word-list
- its frequency in the reference corpus
- the number of running words in the reference corpus

and cross-tabulates these.

Statistical tests include:

- the classic chi-square test of significance with Yates correction for a 2 X 2 table
- Ted Dunning's Log Likelihood test, which gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus.

See UCREL's log likelihood site for more on these.
A word will get into the listing here if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger word-list.

Unusually infrequent key-words are called "negative key-words" and appear at the very end of your listing, in a different colour. Note that negative key-words will be omitted automatically from a keywords database and a plot.

Words which do not occur at all in the reference corpus are treated as if they occurred 5.0e-324 times (0.0000000 and loads more zeroes before a 5) in such a case. This number is so small as not to affect the calculation materially while not crashing the computer's processor.

8.17 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.

The clusters tab below shows us something like this:
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**

### 8.18 KeyWords: links

**The point of it...**

is to find out which key-words are most closely related to a given key-word. A **plot** will show where each key word occurs in the original file. It also shows how many links there are between key-words.

**What are links?**
Links are "co-occurrences of key-words within a collocational span". An example is much easier to understand, though:

Suppose the word *elephant* is key in a text about Africa, and that *water* is also a key word in the same text. If *elephant* and *water* occur within a span of 5 words of each other, they are said to be "linked". The number of times they are linked like this in the text will be shown in the Links window. The link spans (like collocation horizons) go from 1 word away to up to 25 words to left and right. The default 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.

This Links window shows a number of key words followed by the number of linked types (11 here) the total number of hits of the key word (*ROMEO*) and then the individual linked key words. You can if you wish double-click in the *Linked KWs* column and you will see the details listed:
**ROMEO** has 11 linked words; it's linked 23 times with **THOU**, 15 times with **O**, etc. A right-click menu lets you copy or print these details.

**Requirements**

The procedure is text-oriented. You can only get a keywords links list if there is exactly one source text.

Double-click on any word in the **plot listing** to call up a window which show the linked key-words.

See also: Plot calculation, KeyWords clusters, Source Text

---

**8.19 make a word list from keywords data**

With a key word list on your screen, you can press **W** to save your data as a word list (for later comparison, etc. using WordList functions).

**8.20 plot calculation**

**The point of it…**

is to see where the key words are distributed within the text. Do they cluster around the middle or near the beginning of the text?

**How it’s done**

This will calculate the inter-relationships between all the key words identified so far, excluding any which you have deleted or **zapped**.

1. it does a concordance on the text finding all occurrences of each key word;
2. it then works out which of each of the other key words appear within the collocation horizons (set in Settings). It uses the larger of the two horizons.
3. it then plots all the words showing where each occurrence comes in the original file (with a “ruler” showing how many words there are in each part of the file).
4. it computes how many other key-words co-occurred with it, within the current collocational span.
5. it computes a **plot dispersion value**.

Note: this process depends on KeyWords being able to find the **source texts** which your original word-list was based on.

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

See also: Plot Links, Key words plot display

---

**8.21 plot display**

The plot will give you useful visual insights into how often and where the different key words crop up in the text. The plot is initially **sorted** to show which crop up more at the beginning (e.g. in the
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 the listing or use the menu option. The Key word column sorts alphabetically, the dispersion column sorts on the amount of dispersion (higher numbers mean the occurrences are more spread out); the keyness column is the original plot order, or you can sort on number of links with other KWs or on the number of hits found.

**plot data**

You can view the plot data as numbers by double-clicking. Here is the view if one double-clicks on the yellow area:
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 between the key-word and other key-words in the same text, within the current collocation span (default = 5.5). That is, how many times was each key-word found within 5 words of left or right of any of the other key-words in your plot.

**hits**

This column is here to remind you of how many occurrences there were of each key-word. When you have obtained a plot, you can then see the way certain words relate to others. To do this, look at the Links window in the tabs at the bottom, showing which other key words are most linked to the word you clicked on. That is, which other words occur most often within the collocation horizons you've set. The Links window should help you gain insights into the lexical relations here.

Each plot window is dependent on the key words listing from which it was derived. If you close that down, it will disappear. You can Print it. There's no Save option because the plot comes from a key words listing which you should Save, or Save As. There's no save as text option because the plot has graphics, which cannot adequately be represented as text symbols, but you can Copy to the clipboard (Ctrl+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 in the menu (438) allows you to see the plot divided into 8 equal segments if based on one text, or the text-file divisions if there is more than one.

See also: Key words plot, plot dispersion value

### 8.22 regrouping clumps

**How to do it**

You can simply join by dragging, where you think any two clumps belong together because of semantic similarity between their key-words.

Or if you press , KeyWords will inform you which two clumps match best. You'll see a list of the words found only in one, a list of the words found only in the other, and (in the middle) a list of the words which match. It's up to you to judge whether the match is good enough to form a merged
clump. If you aren't sure, press Cancel. If you do want to join them, press Join. If you're sure you don't want to join them and don't want KeyWords to suggest this pair again, press Skip. You can tell KeyWords to skip up to 50 pairs. To clear the memory of the items to be skipped, press Clear Skip.

The point of it (2)…

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

See also: clumps

8.23 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 keyest words appear at the top)

alphabetical order (from A to Z)

certainty in the smaller list (the most frequent words come first)

certainty in the reference list (the most frequent words come first)

A key words plot rotates between sorting by

key-ness (the keyest words appear at the top)

alphabetical order (from A to Z)

frequency (words which appear oftenest come first)

number of links (the most linked words come first)

first mention of each key word in the text

range (words used in smallest sections of text come first)

A key key words database toggles between sorting by

frequency (the most key key words appear at the top)

alphabetical order (from A to Z)

An Associates list toggles between sorting by

frequency (association between title-word and item)

alphabetical order (from A to Z)

frequency (association between item and title-word)
8.24 the key words screen

The display shows:

- each key word
- its frequency in the source text(s) which these key words are key in. (Freq. column below)
- the % that frequency represents
- the number of texts it was present in
- its frequency in the reference corpus (RC. Freq. column)
- the reference corpus frequency as a %
- keyness (chi-square or log likelihood statistic)
- p value
- lemmas (any which have been joined to each other)
- the user-defined set

The calculation of how unusual the frequency is, is based on the statistical procedure used. The statistic appears to the right of the display. If the procedure is log likelihood, or if chi-square is used and the usual conditions for chi-square obtain (expected value >= 5 in all four cells) the probability (p) will be displayed to the right of the chi-square value.

The criterion for what counts as "outstanding" is based on the minimum probability value selected before the key words were calculated. The smaller the number, the fewer key words in the display. Usually you'll not want more than about 40 key words to handle.

The words appear sorted according to how outstanding their frequencies of occurrence are. Those near the top are outstandingly frequent. At the end of the listing you'll find any which are outstandingly infrequent (negative keywords), in a different colour.

There is no upper limit to the keyness column of a set of key words. It is not necessarily sensible to assume that the word with the highest keyness value must be the most outstanding, since keyness is computed merely statistically; there will be cases where several items are obviously equally key (to the human reader) but the one which is found least often in the reference corpus and most often in the text itself will be at the top of the list.
Source text
As its name suggests, choosing the source text tab gets you to a view of the source text(s).

8.25 WordSmith controller: KeyWords settings

These are found in the main Controller 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.

WHAT YOU GET

Procedure
Chi-square or Log Likelihood. The default is Log Likelihood. See procedure for further details.
Max. p value
The default level of significance. See p value for more details.

Max. wanted (500), Min. frequency (3), Min. % of texts (5%)
You may want to restrict the number of key words (KWs) identified so as to find for example the ten most "key" for each text. The program will identify all the key words, sort them by key-ness, and then throw away any excess. It will thus favour positive key words 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 shows the number of texts each KW was found in. (If you see ?? that is because the data were computed before that facility came into WordSmith.)

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

Minimal processing
If this is checked, KeyWords will not compute plots, links or KW clusters as it computes the key words (they can always be computed later assuming you do not move or delete the original text files). This is useful if computing a lot of KW files in a batch, eg. to make a database.

Full lemma processing
If this is checked (the default), KeyWords will compute the full frequency in the case of lemmatised items. For example if GO represents WENT, GOES etc. and GO alone had a frequency of 10 but the whole set GO, WENT, GONE etc. totalled 100, then its frequency will be counted as 100. If unchecked, GO would count only 10.

Max. link frequency
To compute a plot is hard work as all the KWs have to be concordanced so as to work out where they crop up. To compute links between each KW is much harder work again and can take time especially if your KWs include some which occur thousands or hundreds of times in the text. To keep this process more manageable, you can set a default. Here 2000 means that any KW which occurs more than 2000 times in the text will not be used for computing
WHAT YOU SEE

Columns
The *Columns to show/hide* list offers all the standard columns: you may uncheck ones you normally do not wish to see. This will only affect newly computed KeyWords data: earlier data uses the column visibility, size, colours etc already saved. They can be altered using the Layout menu option at any time.

DATABASE

**Database: minimum frequency**
The default is 1. See database.

**Database: associate minimum texts**
The default is 5. See associates.

See also: KeyWords Help Contents, KeyWords calculation.
WordList

Section IX
9 WordList

9.1 purpose

This program generates word lists based on one or more plain text files. The word lists are automatically generated in both alphabetical and frequency order, and optionally you can generate a word index list too.

The point of it...

These can be used

1. simply in order to study the type of vocabulary used;
2. to identify common word clusters;
3. to compare the frequency of a word in different text files or across genres;
4. to compare the frequencies of cognate words or translation equivalents between different languages;
5. to get a concordance of one or more of the words in your list.

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

These word-lists may also be used as input to the KeyWords program, which analyses the words in a given text and compares frequencies with a reference corpus, in order to generate lists of "key-words" and "key-key-words".

Word lists don't have to be of single words, they can be of clusters.

See also: WordList display

Online step-by-step guide showing how

9.2 index

Explanations

What is WordList and what does it do?
Comparing Word-lists
Comparison Display
Consistency Analysis (Simple)
Consistency Analysis (Detailed)
Definitions
Detailed Statistics
Lemmas
9.3 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, 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 from 0.1 to 0.000001 or what you will). The smaller the p value, the more selective the comparison. In other words, a p setting of 0.1 will show more words than a p setting of 0.0001 will.

The display format is similar to that used in KeyWords. You will also find the Dice coefficient which compares the vocabularies of the two texts, reported in the Notes.

See also: Compute Key Words, Consistency Analysis, Match List

9.4 merging wordlists

The point of it
You might want to merge 2 word lists (or concordances, mutual information lists etc.) with each other if making each one takes ages or if you are gradually building up a master word list or concordance based on a number of separate genres or text-types.

How to do it
With one word-list (or concordance) opened, choose File | Merge with and select another.

Be aware that...
Making a merged word list implies that each set of source texts was different. If you choose to merge 2 word lists both of which contained information about the same text file, WordSmith will do as you ask even though the information about the number of occurrences and of texts in which each word-type was found is (presumably) inaccurate.

Merging a list in English with another in Spanish: if you start with the one in Spanish, the one in English will be merged in and henceforth treated as if it were Spanish, eg. in sort order. Presumably if you try to merge one in English with one in Arabic (I've never tried) you should see all the forms but you would get different results merging the Arabic one into the English one (all the Arabic words would be treated as if they were English).
9.5 comparison display

How to get here? by choosing compare two wordlists

Here is a comparison window, where we have compared Shakespeare's King Lear with Romeo and Juliet.

The display shows

- frequency in the text you started with, here King Lear, (with % if > 0.01%) -- then, to the right
- frequency in the other text, here Romeo & Juliet, (with % if > 0.01%) -- then, to the right
- chi-square or log likelihood, and p value,

The criterion for what counts as "outstanding" is based on the minimum probability value entered before the lists were compared. The smaller this probability value the fewer words in the display.

The words appear sorted according to how outstanding their frequencies of occurrence are. Those near the top are outstandingly frequent in your main 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 program. The KeyWords analysis is slightly quicker and allows for batch processing.

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

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

<table>
<thead>
<tr>
<th>N</th>
<th>Key word</th>
<th>Freq.</th>
<th>2</th>
<th>% Freq.</th>
<th>RC.</th>
<th>2</th>
<th>% Keyness</th>
<th>P</th>
<th>emwad</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>DAUGHTERS</td>
<td>28</td>
<td>0.10</td>
<td>2</td>
<td>25.34</td>
<td>01786</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>TOM</td>
<td>19</td>
<td>0.07</td>
<td>0</td>
<td>25.20</td>
<td>05143</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>HAVE</td>
<td>207</td>
<td>0.75</td>
<td>127</td>
<td>0.49</td>
<td>24.72</td>
<td>06566</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>MOST</td>
<td>54</td>
<td>0.20</td>
<td>16</td>
<td>0.06</td>
<td>24.21</td>
<td>08868</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>DEATH</td>
<td>22</td>
<td>0.08</td>
<td>71</td>
<td>0.27</td>
<td>24.29</td>
<td>08253</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>EALTHASAR</td>
<td>0</td>
<td>0.00</td>
<td>19</td>
<td>0.07</td>
<td>25.20</td>
<td>05143</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>GREGORY</td>
<td>0</td>
<td>0.00</td>
<td>19</td>
<td>0.07</td>
<td>25.20</td>
<td>05143</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>MARRIED</td>
<td>0</td>
<td>0.00</td>
<td>19</td>
<td>0.07</td>
<td>25.20</td>
<td>05143</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

### 9.6 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 with others in WordList, or using KeyWords. This way you can determine which are the significantly consistent words in your genre or sub-genre.

See also: Consistency Analysis (Detailed), Comparing Word-lists, Match List

9.7 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 defined in the settings 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, word-list with tags as prefix

9.8 detailed consistency analysis

This function does exactly the same thing as simple consistency, but provides much more detail.

The point of it...

The idea is to help stylistic comparisons. Suppose you're studying several versions of a story, or different translations of it. This function enables you to see all the words which are used in the word lists which you have called up.
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 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.

**How to do it...**

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

**Settings**
You can require a minimum number of texts and minimum frequency in the main **Controller** if you click this.

**Sorting**
Each column can be sorted by clicking on its header column (**Word**, **Freq.** etc.). To get the words which occurred in all 37 texts to the top, I clicked **Texts**.

**Row percentages**
If you choose to **Show as %**, 
you will transform the view so as to get row percentages. In this screenshot, we see the last few items which appear only in Anthony and Cleopatra, then Cleopatra (93.3%), Egypt (93.18%) etc. (Egypt appears also in A Midsummer Night's Dream, As You Like It, King Henry VIII.)
See also: Detailed Consistency Relations, Consistency Analysis (range), Comparison Display, Comparing Word-lists, Match List, Column Totals.

9.9 detailed consistency relations

With a detailed consistency list such as this, of five versions of the fairy story *Little Red Riding Hood*,

<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Total</th>
<th>Texts</th>
<th>as Set</th>
<th>red1</th>
<th>red2</th>
<th>red3</th>
<th>red4</th>
<th>red5</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>HER</td>
<td>100</td>
<td>5</td>
<td>0</td>
<td>9</td>
<td>22</td>
<td>25</td>
<td>6</td>
<td>38</td>
</tr>
<tr>
<td>5</td>
<td>SHE</td>
<td>96</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>14</td>
<td>20</td>
<td>5</td>
<td>62</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>63</td>
<td>5</td>
<td>0</td>
<td>12</td>
<td>12</td>
<td>15</td>
<td>9</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>YOU</td>
<td>70</td>
<td>5</td>
<td>0</td>
<td>14</td>
<td>15</td>
<td>22</td>
<td>1</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>RED</td>
<td>66</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>20</td>
<td>1</td>
<td>31</td>
</tr>
<tr>
<td>9</td>
<td>OF</td>
<td>66</td>
<td>5</td>
<td>0</td>
<td>14</td>
<td>8</td>
<td>9</td>
<td>4</td>
<td>32</td>
</tr>
<tr>
<td>10</td>
<td>WAS</td>
<td>56</td>
<td>5</td>
<td>0</td>
<td>3</td>
<td>9</td>
<td>9</td>
<td>4</td>
<td>31</td>
</tr>
<tr>
<td>11</td>
<td>GRANDMOTHER</td>
<td>53</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>17</td>
<td>5</td>
<td>19</td>
</tr>
<tr>
<td>12</td>
<td>LITTLE</td>
<td>51</td>
<td>5</td>
<td>0</td>
<td>7</td>
<td>15</td>
<td>17</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>IN</td>
<td>47</td>
<td>5</td>
<td>0</td>
<td>5</td>
<td>7</td>
<td>10</td>
<td>3</td>
<td>22</td>
</tr>
<tr>
<td>14</td>
<td>HOOD</td>
<td>44</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td>28</td>
</tr>
</tbody>
</table>

it looks as if the most long-winded story is probably version 5 (*red*5.lst). If you click the detailed cons. relation tab

![Detailed consistency relations tab](image)

you can see the relevant statistics more usefully:

<table>
<thead>
<tr>
<th>N</th>
<th>File 1</th>
<th>Count</th>
<th>File 2</th>
<th>Count</th>
<th>Joint</th>
<th>Relation</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>red1.lst</td>
<td>169</td>
<td>red2.lst</td>
<td>234</td>
<td>83</td>
<td>0.412</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>red1.lst</td>
<td>169</td>
<td>red3.lst</td>
<td>333</td>
<td>69</td>
<td>0.355</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>red1.lst</td>
<td>169</td>
<td>red4.lst</td>
<td>98</td>
<td>42</td>
<td>0.315</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>red1.lst</td>
<td>169</td>
<td>red5.lst</td>
<td>462</td>
<td>88</td>
<td>0.282</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>red2.lst</td>
<td>234</td>
<td>red3.lst</td>
<td>333</td>
<td>138</td>
<td>0.437</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>red2.lst</td>
<td>234</td>
<td>red4.lst</td>
<td>98</td>
<td>57</td>
<td>0.343</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>red2.lst</td>
<td>234</td>
<td>red5.lst</td>
<td>462</td>
<td>136</td>
<td>0.391</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>red3.lst</td>
<td>333</td>
<td>red4.lst</td>
<td>98</td>
<td>61</td>
<td>0.283</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>red3.lst</td>
<td>333</td>
<td>red5.lst</td>
<td>462</td>
<td>162</td>
<td>0.403</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>red4.lst</td>
<td>98</td>
<td>red5.lst</td>
<td>462</td>
<td>61</td>
<td>0.218</td>
<td></td>
</tr>
</tbody>
</table>

where it can be seen that *red*5 has a word-count of 462 words, more than any other, and that the relation between *red*2 and *red*3 is the closest with a relation statistic of 0.487. This relation is the Dice coefficient, based on the joint frequency (there are 138 matches in the vocabulary of these
two versions) and the word-counts of the two texts. 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

9.10 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.

```
<table>
<thead>
<tr>
<th>N</th>
<th>Word</th>
<th>Freq</th>
<th>% Texts</th>
<th>% Emerges</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,629</td>
<td>EVERY ONE DOTH</td>
<td>3</td>
<td>4</td>
<td>10.01</td>
<td></td>
</tr>
<tr>
<td>1,630</td>
<td>EVERY THING IS</td>
<td>5</td>
<td>3</td>
<td>8.11</td>
<td></td>
</tr>
<tr>
<td>1,631</td>
<td>EYES AND EARS</td>
<td>8</td>
<td>7</td>
<td>18.92</td>
<td></td>
</tr>
<tr>
<td>1,632</td>
<td>EYES TO SEE</td>
<td>5</td>
<td>5</td>
<td>13.51</td>
<td></td>
</tr>
<tr>
<td>1,633</td>
<td>FACE TO FACE</td>
<td>5</td>
<td>5</td>
<td>13.51</td>
<td></td>
</tr>
<tr>
<td>1,634</td>
<td>FAIN WOULD I</td>
<td>5</td>
<td>5</td>
<td>13.51</td>
<td></td>
</tr>
<tr>
<td>1,637</td>
<td>FAITH I AM</td>
<td>6</td>
<td>6</td>
<td>16.72</td>
<td></td>
</tr>
</tbody>
</table>
```

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:
9.11 Lemmas (joining words)

9.11.1 what are lemmas and how do we join words?

In a word list, a key word list or a list of collocates you may want to store several entries together: e.g. want; wants; wanting; wanted. Bringing them together means you're treating them as members of the same "lemma" or set -- rather like a headword in a dictionary.

A lemmatised head entry has a red mark in the left margin beside it. The others you marked will be coloured as if deleted. The linked entries which have been joined to the head can be seen at the right.

Here we see a word list based on 3-word clusters where originally a good deal had a frequency of 24, but has been joined to a great deal and a good few and thereby risen to 141.

Joining can be done automatically or manually.

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.

**Get rid of the deleted words**

If you don't want to see the deleted words

<table>
<thead>
<tr>
<th>#</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>33,32</td>
<td>1.18</td>
<td>430</td>
<td>89.58</td>
<td>32,24</td>
<td>4.14</td>
<td>480</td>
<td>100.04</td>
<td>28,43</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>27,43</td>
<td>0.97</td>
<td>431</td>
<td>89.78</td>
<td>25,02</td>
<td>0.88</td>
<td>480</td>
<td>100.04</td>
<td>21,79</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>21,34</td>
<td>0.75</td>
<td>475</td>
<td>98.96</td>
<td>21,33</td>
<td>0.75</td>
<td>416</td>
<td>86.67</td>
<td>85,58</td>
<td>3.02</td>
</tr>
<tr>
<td></td>
<td>17,68</td>
<td>0.63</td>
<td>480</td>
<td>100.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

choose Ctrl-Z to **zap** them.

See also: Auto-Joining methods, Using a text file to lemmatise, selecting multiple entries, Concord lemmatisation.

**9.11.2 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 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.

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 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.
9.11.3 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 which automates the matching & joining process. The actual processing of the list takes place when you choose the menu option Match Lemmas (Ξ) 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 *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 (Ctrl+F4) any mistakes than to confirm each one... Finally, sort on the Word and save.

See also: Lemmatisation
9.11.4 choosing lemma file

The point of it...
You may choose to lemmatise all items in the current word-list using a standard text file which groups words which belong together (be -> was, is, were, etc.). While it is time-consuming producing the text file the first time, it will be very useful if you want to lemmatise lots of word lists, and is much less "hit-and-miss" than auto-joining using a template.

There is an English-language lemma list from Yasumasa Someya at http://lexically.net/downloads/BNC_wordlists/e_lemma.txt.

How to do it
Lemma list settings are accessed via the Lists option in the WordList menu or an Advanced Settings button in the Controller.
Choose the appropriate button (for Concord, KeyWords or WordList) and type the file name or browse for it, then Load it.

The file should contain a plain text list of lemmas with items like this:

```
BE  ->  AM, ARE, WAS, WERE, IS
GO  ->  GOES, GOING, GONE, WENT
```

WordSmith then reads the file and displays them (or a sample if the list is long). The format allows any alphabetic or numerical characters in the language the list is for, plus the single apostrophe, space, underscore. In other words, if you mistakenly put `GO = GOES` that line won't be included because of the = symbol.

The actual processing of the list will take place when you compute your word list, key word list or concordance or when you choose the menu option `Match Lemmas` in WordList, Concord or KeyWords. See `Match List` for a more detailed explanation, with screenshots. Lemmatising occurs before any `stop list` is processed.

**What if my text files don’t contain the headword of the lemma?**

Suppose you are matching `AM`, `ARE` etc with `BE` as in the list above, but your texts don't actually contain the word `BE`. In that case the tool will insert `BE` with zero frequency and add `AM`, `ARE` etc as needed.

See also: `Lemmatisation`, `Match List`, `Stop List`, `Lemmatisation in Concord`.

### 9.12 WordList Index

#### 9.12.1 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 

2. Another is to compute `Mutual Information` scores which relate word types to each other.
3. Or you can use an index to see word clusters.
4. Finally, an index is needed to generate concgram searches.

See also Making an Index List, Viewing Index Lists, Exporting index data, find filenames for word clusters, WordList Help Contents, WSConcgram.

9.12.2 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 and Mutual Information scores for each word type. The screenshot below shows the progress bars for an index of the BNC corpus; on a modern PC it might work at a rate of about 2.8 million words per minute. The resulting BNC.tokens file was 1.6GB in size and the BNC.types file was 26 MB.

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.
9.12.3 index clusters

WordList clusters

A word list doesn't need to be of single words. You can ask for a word list consisting of two, three, up to eight words on each line. To do cluster processing in WordList, first make an index.

How to see clusters...

Open the index. Now choose Compute | Clusters.

To choose words which aren’t next to each other, press Control and click in the number at the left -- keep Control held down and click elsewhere. The first one clicked will go green and the others white. In the picture below, using an index of the BNC corpus, I selected world and then life by clicking numbers 164 and 167.

- "all": all the clusters involving all words above a certain frequency (this will be s-l-o-w for a big corpus like the BNC), or
- "selection": clusters only for words you've selected (eg. you have highlighted BOOK and BOOKS and you want clusters like book a table, in my book).

To choose words which aren't next to each other, press Control and click in the number at the left -- keep Control held down and click elsewhere. The first one clicked will go green and the others white. In the picture below, using an index of the BNC corpus, I selected world and then life by clicking numbers 164 and 167.
The process will take time. In the case of BNC, the index knows the positions of all of the 100 million words. To find 3-word clusters, in the case above, it took about a minute to process all the 115,000 cases of world and life and find 5,719 clusters like the world bank and of real life. Chris Tribble tells me it took his PC 36 hours to compute all 3-word clusters on the whole BNC ... he was able to use the PC in the meantime but that's not a job you're going to want to do often.

- What you see

The cluster size must be between 2 and 8 words.
The min. frequency is the minimum number of each that you want to see.
 omit #: if selected, this won't show any clusters involving numbers and dates
 omit phrase frames: see phrase frames section below.
Here the user has chosen to see any 3-4-word clusters that appear 5 or more times.

- Working constraints

The "max. frequency %" setting is to speed the process up.
- in more detail...

It means the maximum frequency percentage which the calculation of clusters for a given word will process. This is because there are lots and lots of the very high frequency items and you may well not be interested in clusters which begin with them. For example, the item the is likely to be about 6% of any word-list (about 6 million of them in the BNC therefore), and you might not want clusters starting the... -- if so, you might set the max. percent to 0.5% or
0.1% (which for the BNC corpus will cut out the top 102 frequency words). You will still get clusters which include very high frequency items in the middle or end, like the a in book a table, but would not get in my book, which begins with the very high frequency word in. The more words you include, the longer the process will take....

*Stop at*, like Concord clusters, offers a number of constraints, such as *sentence and other punctuation-marked breaks*. The idea is that a 5-word cluster which starts in one sentence and continues in the next is not likely to make much sense.

*Max. seconds per word* is another way of controlling how long the process will take. The default (0) means no limit. But if you set this e.g. to 30 then as WordList processes the words in order, as soon as one has taken 30 seconds no further clusters will be collected starting with that word.

*batch processing* allows you to create a whole set of cluster word-lists at one time.

**Phrase frames**

These are what William H. Fletcher has defined as *phrase-frames*, i.e. "groups of wordgrams identical but for a single word", in his kfNgram program.

Here, processing 23 Dickens novels shows lots of phrase frames where the wildcard word is represented with *.

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).

**Omit phrase frames?**

If you don't want to see phrase frames, select the *omit phrase frames* option.

![Cluster choices](image1)

Here below, the listing has all *his hand* sequences together but not *drawing his hand across, gave his hand to*, etc. as shown in the phrase frame view above.

![Index file: Dickens](image2)
Here is a small set of 3-word clusters involving rabies from the BNC corpus.

It's a word list

Finally, remember this listing is just like a single-word word list. You can save it as a .lst file and open it again at any time, separately from the index.

See also: find the files for specific clusters, clusters in Concord

9.12.4 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 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*.

**9.12.5 index lists: viewing**

In WordList, open an index as you would any other kind of word-list file -- using File | Open. The filename will end `.tokens`. Easier, in the Controller | Previous lists, choose any index you've made and double-click it.

The index *looks* exactly like a large word-list. (Underneath, it "knows" a lot more and can do more but it looks the same.)
The picture above shows the top 10 words in the BNC Corpus. Number 5 (#) represents numbers or words which contain numbers such as £50.00. These very frequent words are also very consistent – they appear in at least 99% of the 4,054 texts of BNC.

In the view below, you see words sorted by the number of Texts: all these words appeared 10 times in the corpus but their frequencies vary.

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 like these:
9.12.6 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 will be quicker if you leave the word details space empty.

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

9.14.1 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.

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

There are various different formulae for computing the strength of collocational relationships. The MI in WordSmith ("specific mutual information") is computed using a formula derived from Gaussier, Lange and Meunier described in Oakes, p. 174; here the probability is based on total corpus size in tokens. Other measures of collocational relation are computed too, which you will see explained under Mutual Information Display.

Settings

The Relationships settings are found in the Controller under Main Settings | Advanced | Index or in a menu option in WordList.

See also: Mutual Information Display, Computing Mutual Information, Making an Index List, Viewing Index Lists, WordList Help Contents.

See Oakes for further information about Mutual Information, Dice, MI3 etc.

9.14.2 relationships display

The Relationships procedure contains a number of columns and uses various formulae.
**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", 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 (not just the joint frequency at the typical gap distance).

In line 7 of this display, **BACKWARDS** occurs 83 times in the whole index (based on Dickens novels), and **FORWARDS** 8 times. They occur together 62 times. The gap is 2 because **backwards**, in these data, typically comes 2 words away from **forwards**. The pair **backwards * forwards** comes in 17 texts. (This search was computed using the **to right only** setting mentioned above).

As usual, the data can be sorted by clicking on the headers. Let's now sort by clicking on "Z score" first and "Word 1" second.

You get a double sort, main and secondary, because sometimes you will want to see how MI or Z score or other sorting affects the whole list and sometimes you will want to keep the words sorted alphabetically and only sort by MI or Z score within each word-type. Press **Swap** to switch the primary & secondary sorts.
The order is not quite the same ... but not very different either. Both Freq. columns have fairly small numbers.

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

<table>
<thead>
<tr>
<th>N</th>
<th>Word 1</th>
<th>Freq.</th>
<th>Word 2</th>
<th>Freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AIDER</td>
<td>5</td>
<td>ABETTOR</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>BACKWARDS</td>
<td>83</td>
<td>FORWARDS</td>
<td>67</td>
</tr>
<tr>
<td>3</td>
<td>BEG</td>
<td>639</td>
<td>PARDON</td>
<td>544</td>
</tr>
<tr>
<td>4</td>
<td>ANIMATE</td>
<td>14</td>
<td>INANIMATE</td>
<td>20</td>
</tr>
<tr>
<td>5</td>
<td>BEAN</td>
<td>9</td>
<td>STALK</td>
<td>19</td>
</tr>
<tr>
<td>6</td>
<td>BEAU</td>
<td>27</td>
<td>Ti</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>BEAU</td>
<td>27</td>
<td>FUL</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>BLINDMAN'S</td>
<td>5</td>
<td>BUFF</td>
<td>43</td>
</tr>
<tr>
<td>9</td>
<td>BOW</td>
<td>253</td>
<td>WOW</td>
<td>28</td>
</tr>
<tr>
<td>10</td>
<td>AREA</td>
<td>87</td>
<td>RAILINGS</td>
<td>44</td>
</tr>
<tr>
<td>11</td>
<td>BREAD</td>
<td>370</td>
<td>BUTTER</td>
<td>168</td>
</tr>
</tbody>
</table>
Much more frequent items have jumped to the top.

Now, by Log Likelihood (Dunning, 1993):

Here the Word 2 items are again very high frequency ones and we get at colligation (grammatical collocation). A T Score listing is fairly similar:
but a Dice score ordered list brings us back to results akin to the first two shown above:

See also: Formulae, Mutual Information and other relationships, Computing Relationships, Making an Index List, Viewing Index Lists, WordList Help Contents.

See Oakes for further information about the various statistics offered.
9.14.3 relationships computing

To compute these relationship statistics you need a WordList Index. 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 (with a). Or you can always do part of the list (eg. A to D) and later merge your mutual-information list with another (E to H).

  Alternatively you may choose to use only items from a plain text file constructed using the same syntax as a match-list file., or to use all items except ones from your plain text file.

- **omissions**

  *omit any containing #* will cut out numbers, and *omit if word1=word2* is there because you might
find that **GOOD** is related to **GOOD** if there are lots of cases where these 2 are found near each other.

**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**:

<table>
<thead>
<tr>
<th>N</th>
<th>Word 1</th>
<th>Freq</th>
<th>Word 2</th>
<th>Freq</th>
<th>Texts</th>
<th>Gap</th>
<th>Joint</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DUST</td>
<td>69</td>
<td>WITH</td>
<td>7,143</td>
<td>9</td>
<td>1</td>
<td>11</td>
<td>4.41</td>
</tr>
<tr>
<td>2</td>
<td>VITH</td>
<td>7,148</td>
<td>DUST</td>
<td>69</td>
<td>9</td>
<td>1</td>
<td>11</td>
<td>4.41</td>
</tr>
</tbody>
</table>

**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.

<table>
<thead>
<tr>
<th>N</th>
<th>Word 1</th>
<th>Freq</th>
<th>Word 2</th>
<th>Freq</th>
<th>Texts</th>
<th>Gap</th>
<th>Joint</th>
<th>MI</th>
<th>Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DUST</td>
<td>53</td>
<td>WITH</td>
<td>7,148</td>
<td>5</td>
<td>1</td>
<td>6</td>
<td>3.84</td>
<td>0.37</td>
</tr>
</tbody>
</table>

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 determine that one word cannot be related to another. With stop at sentence break, “I wrote the letter”.

---

*February, 2015. Page 287*
Then I posted it" would not consider posted as a possible collocate of letter because there's a sentence break between them.

spars: 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: sometimes years ago it took over an hour to compute MI for all words beginning with B in the case of the BNC edition (written, 90 million words) in the screenshot below, using default settings. It might take 24 hours to process the whole BNC, 100 million words, even on a modern powerful PC. Don't forget to save your results afterwards!

<table>
<thead>
<tr>
<th>Word 1</th>
<th>Freq</th>
<th>Word 2</th>
<th>Freq</th>
<th>Texts</th>
<th>Gap</th>
<th>Joint</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>15,642</td>
<td>BUDGETS</td>
<td>1,171</td>
<td>CONTRACTS</td>
<td>4,327</td>
<td>4</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>19,643</td>
<td>BUDIMIR</td>
<td>13</td>
<td>LONCAR</td>
<td>17</td>
<td>8</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>19,644</td>
<td>BUDS</td>
<td>404</td>
<td>STAR</td>
<td>8,817</td>
<td>3</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>19,645</td>
<td>BUDS</td>
<td>404</td>
<td>FLOWERS</td>
<td>5,038</td>
<td>5</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>19,646</td>
<td>BUDS</td>
<td>404</td>
<td>FRUIT</td>
<td>3,799</td>
<td>4</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>19,647</td>
<td>BUDS</td>
<td>404</td>
<td>SHOOTS</td>
<td>567</td>
<td>6</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>19,648</td>
<td>BUENA</td>
<td>9</td>
<td>VISTA</td>
<td>158</td>
<td>4</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>19,649</td>
<td>BUENOS</td>
<td>6</td>
<td>NOCHES</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>19,650</td>
<td>BUENOS</td>
<td>210</td>
<td>MARAVI</td>
<td>15,665</td>
<td>4</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>19,651</td>
<td>BUENOS</td>
<td>210</td>
<td>ARES</td>
<td>201</td>
<td>94</td>
<td>1</td>
<td>176</td>
</tr>
<tr>
<td>19,652</td>
<td>BUFF</td>
<td>252</td>
<td>WHITE</td>
<td>23,765</td>
<td>3</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>19,653</td>
<td>BUFF</td>
<td>252</td>
<td>COLOURED</td>
<td>3,285</td>
<td>13</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>19,654</td>
<td>BUFF</td>
<td>252</td>
<td>BROWN</td>
<td>8,238</td>
<td>3</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>19,655</td>
<td>BUFF</td>
<td>252</td>
<td>ENVELOPE</td>
<td>1,163</td>
<td>6</td>
<td>1</td>
<td>9</td>
</tr>
<tr>
<td>19,656</td>
<td>BUFFALO</td>
<td>307</td>
<td>YORK</td>
<td>7,689</td>
<td>5</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td>19,657</td>
<td>BUFFALO</td>
<td>307</td>
<td>TOM</td>
<td>6,898</td>
<td>4</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>19,658</td>
<td>BILL</td>
<td>12,164</td>
<td>7</td>
<td>1</td>
<td>9</td>
<td>7.73</td>
<td>13.17</td>
</tr>
<tr>
<td>19,659</td>
<td>BILL</td>
<td>2020</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>9.67</td>
<td>25.23</td>
</tr>
</tbody>
</table>

See also Collocates, Mutual Information Settings, Mutual Information Display, Detailed Consistency Relations, Making an Index List, Viewing Index Lists, Recompute Token Count, WordList Help Contents.

9.15 recompute tokens

Why recompute the tokens?

To compute relations such as Mutual Information or Keyness 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, apostrophes and other non-letters, in the middle of a word, numbers, words cut out because of a stoplist 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?

For most normal purposes, the total number of running words (tokens) computed when the word list or index was created will be used for these statistical calculations.

**How to do it**

Compute | Tokens

![Compute Tokens](compute_tokens.png)

**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.

### 9.16 re-sorting: consistency lists

The frequency-ordered consistency display can be re-sorted by alphabetical order (Word) total frequencies overall (Total, the default) by the frequencies in any given file (you see the file names).

Click on Word, Total or a filename to choose.

The sort can be either ascending or descending, the default being descending.

See also: Sorting word-lists

### 9.17 statistics

#### 9.17.1 statistics

Visible by clicking the Statistics tab at the bottom of a WordList window:
Overall results take up the top row. Details for the individual text files follow below.

Statistics include:
- number of files involved in the word-list
- file size (in bytes, i.e. characters)
running words in the text (*tokens*)
tokens used in the list (would be affected by using a *stoplist* or changes to *minimum settings*)

sum of entries: choose *Compute / Tokens* to see, otherwise this will be blank

no. of different words (*types*)
type/token ratios

no. of *sentences* in the text
mean sentence length (in words)
standard deviation of sentence length (in words)

no. of *paragraphs* in the text
mean paragraph length (in words)
standard deviation of paragraph length (in words)

no. of *headings* in the text (none here because WordSmith didn't know how to recognise headings)
mean heading length (in words)
standard deviation of heading length (in words)

*numbers* removed
*stoplist* tokens and types removed

the number of *n*-letter words (to see these scroll the grid horizontally)

(14 is the *default* maximum word length. But you can set it to any length up to 50 letters in Word List Settings, in the Settings menu.) Longer words are cut short but this is indicated with a + at the end of the word.

The number of *types* (different words) is computed separately for each text. Therefore if you have done a single word-list involving more than one text, summing the number of types for each text will not give the same total as the number of types over the whole collection.

**Vertical layout**

If you prefer the layout in previous versions of WordSmith, you can choose to save the statistics vertically in a text file.
This lets you choose which ones (any unchecked are zero in the data):

and the data will be saved listed vertically.
Alternatively you could export the data here to Excel and use its Transpose function to get the rows and columns swapped.
Tokens used for word list

In these data, there were over 2.8 million running words of text, but 38,943 numbers were not listed separately, so the number of tokens in the word-list is a little under 2.8 million.

<table>
<thead>
<tr>
<th>Numbers removed</th>
<th>Stoplist tokens removed</th>
<th>Stoplist types removed</th>
<th>1-letter words</th>
<th>2-letter words</th>
<th>3-letter words</th>
<th>4-letter words</th>
</tr>
</thead>
<tbody>
<tr>
<td>38,943</td>
<td>108,75</td>
<td>499,84</td>
<td>581,67</td>
<td>485,59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>187</td>
<td>257</td>
<td>1,133</td>
<td>1,258</td>
<td>1,273</td>
<td></td>
<td></td>
</tr>
<tr>
<td>183</td>
<td>322</td>
<td>1,444</td>
<td>1,683</td>
<td>1,618</td>
<td></td>
<td></td>
</tr>
<tr>
<td>104</td>
<td>205</td>
<td>995</td>
<td>1,203</td>
<td>1,036</td>
<td></td>
<td></td>
</tr>
<tr>
<td>221</td>
<td>284</td>
<td>1,660</td>
<td>2,266</td>
<td>2,012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>92</td>
<td>148</td>
<td>717</td>
<td>929</td>
<td>777</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>123</td>
<td>649</td>
<td>816</td>
<td>600</td>
<td></td>
<td></td>
</tr>
<tr>
<td>284</td>
<td>288</td>
<td>1,747</td>
<td>1,771</td>
<td>1,440</td>
<td></td>
<td></td>
</tr>
<tr>
<td>110</td>
<td>227</td>
<td>1,574</td>
<td>1,720</td>
<td>1,338</td>
<td></td>
<td></td>
</tr>
<tr>
<td>172</td>
<td>255</td>
<td>1,570</td>
<td>1,904</td>
<td>1,487</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>117</td>
<td>854</td>
<td>896</td>
<td>736</td>
<td></td>
<td></td>
</tr>
<tr>
<td>475</td>
<td>363</td>
<td>1,986</td>
<td>2,114</td>
<td>1,628</td>
<td></td>
<td></td>
</tr>
<tr>
<td>78</td>
<td>247</td>
<td>1,657</td>
<td>1,698</td>
<td>1,318</td>
<td></td>
<td></td>
</tr>
<tr>
<td>276</td>
<td>282</td>
<td>1,362</td>
<td>1,616</td>
<td>1,370</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**MS Word's word count is different!**

The number of tokens found is affected by your settings such as treatment of numbers, hyphens and mid-word letter settings (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 (with a screenshot), Summary Statistics, Starts and Ends of Text Segments, Recomputing tokens.
9.17.2 summary statistics

A word list's statistics give you data about the corpus, but you may need more specific information about individual words in a word list too.

**How many end in -ly?**

1. type a search (using Concord syntax)
2. choose the relevant column of data

Press Count

![Image of software interface]

...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

**Load button**
see the explanation for count data frequencies

### 9.18 stop-lists and match-lists

In WordList, a stop list 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 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.

See also: lemma lists, general explanation of stop-lists

9.19 import words from text list

the point of it
You might want a word list based on some data you have obtained in the form of a list, but whose original texts you do not have access to.

requirements
Your text file can be in any language (select this before you make the list), and can be in Unicode or ASCII or ANSI, plain text. <Tab> characters are expected to separate the columns of data. Decimal points and commas will be ignored. 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  67543
IT    33218
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*, 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*.

### 9.20 type/token ratios

If a text is 1,000 words long, it is said to have 1,000 "tokens". But a lot of these words will be repeated, and there may be only say 400 different words in the text. "Types", therefore, are the different words.

The ratio between types and tokens in this example would be 40%.

But this type/token ratio (TTR) varies very widely in accordance with the length of the text -- or corpus of texts -- which is being studied. A 1,000 word article might have a TTR of 40%; a shorter
one might reach 70%; 4 million words will probably give a type/token ratio of about 2%, and so on. Such type/token information is rather meaningless in most cases, though it is supplied in a WordList statistics display. The conventional TTR is informative, of course, if you're dealing with a corpus comprising lots of equal-sized text segments (e.g. the LOB and Brown corpora). But in the real world, especially if your research focus is the text as opposed to the language, you will probably be dealing with texts of different lengths and the conventional TTR will not help you much.

WordList 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, \( n = 1,000 \). In other words the ratio is calculated for the first 1,000 running words, then calculated afresh for the next 1,000, and so on to the end of your text or corpus. A running average is computed, which means that you get an average type/token ratio based on consecutive 1,000-word chunks of text. (Texts with less than 1,000 words (or whatever \( n \) is set to) will get a standardised type/token ratio of 0.)

### Setting the N boundary
Adjust the \( n \) number in Minimum & Maximum Settings to any number between 100 and 20,000.

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

The number shown is a percentage of new types for every \( n \) tokens. That way you can compare type/token ratios across texts of differing lengths. This method contrasts with that of Tuldava (1995:131-50) who relies on a notion of 3 stages of accumulation. The WordSmith method of computing STTR was my own invention but parallels one of the methods devised by the mathematician David Malvern working with Brian Richards (University of Reading).

### Further discussion
TTR and STTR are both pretty crude measures even if they are often assumed to imply something about "lexical density". Suppose you had a text which spent 1,000 words discussing ELEPHANT, LION, TIGER etc, and then 1,000 discussing MADONNA, ELVIS etc, then 1,000 discussing CLOUD, RAIN, SUNSHINE. If you set the STTR boundary at 1,000 and happened to get say 48% or so for each section, the statistic in itself would not tell you there was a change involving Africa, Music, Weather. Suppose the boundary between Africa & Music came at word 650 instead of at word 1,000, I guess there'd be little or no difference in the statistic. But what would make a difference? A text which discussed clouds and written by a person who distinguished a lot between types of cloud might also use MIST, FOG, CUMULUS, CUMULO-NIMBUS. This would be higher in STTR than one written by a child who kept referring to CLOUD but used adjectives like HIGH, LOW, HEAVY, DARK, THIN, VERY THIN to describe the clouds... and who repeated DARK, THIN, etc a lot in describing them......

(NB. Shakespeare is well known to have used a rather limited vocabulary in terms of measures like these!)
9.21 case sensitivity

Normally, you'll make a case-insensitive word list, especially as in most languages capital letters are used not only to distinguish proper nouns but also to signal beginnings of sentences, headings, etc. If, however, you wish to make a word list which distinguishes between major, Major and MAJOR, activate case sensitivity ([WordList Settings | Case Sensitivity in the Controller](#)).

When you first see your case-sensitive list, it is likely to appear all in UPPER CASE. Press Ctrl+L or choose the Layout menu option to change this.

9.22 minimum & maximum settings

These include:

- **minimum word length**
  Default: 1 letter. When making a word-list, you can specify a minimum word length, e.g. so as to cut out all words of less than 3 letters.

- **maximum word length**
  Default: 49 letters. You can allow for words of up to 50 characters in length. If a word exceeds the limit and Abbreviate with + is checked, WordList will append a + symbol at the end of it to show that it was cut short. (If Abbreviate with + is not checked, the long word will be omitted from your word list. You might wish to use this to set both minimum and maximum to say, 4, and leave Abbreviate with + un-checked – that way you'll get a word-list with only the 4-letter words in it.

- **minimum frequency**
  Default: 1. By default, all words will be stored, even those which occur once only. If you want only the more frequent words, set this to any number up to 32,000.

- **maximum frequency**
  Default maximum is 2,147,483,647 (2 Gigabytes). You'd have to analyse a lot of text to get a word which occurred as frequently as that!. You might set this to say 500, and the minimum to 50: that way your word-list would hold only the moderately common words.

- **type/token mean number (default 1,000)**
  Enables a smoothed calculation of type/token ratio for word lists. Choose a number between 10 and 20,000. For a more complete explanation, see [WordList Type/Token Information](#).

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

9.23 sort order

**How to do it...**
Sorting can be done simply by pressing the top row of any list. Press again to toggle between ascending & descending sorts.
With a word-list on your screen, the main Frequency window doesn’t sort, but you can re-sort the Alphabetical window (look at the tabs at the bottom of WordList to choose the tab) in a number of different ways.

The menu offers various options.

**Alphabetical Word Sort**

Many languages have their own special sorting order, so prior to sorting or re-sorting, check that you have selected the right language for the words being sorted. Spanish, for example, uses this order: A, B, C, CH, D, E, F, G, H, I, J, K, 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.

See also:  Concord sort, KeyWords sort, Editing entries, Accented characters, Choosing Language

### 9.24 WordList and tags

If you have defined a tag file and made the appropriate settings to load it, you can get a word-list which treats tags and words separately as in this example, where the tag is viewed as if it were a prefix.

#### A word list only of tags?

Choose whether you want only the tags, only the words or both in WordList settings | What you see | Tags:

- **words only, no tags**
- **tags as “prefix” to words**
- **tags only, no words**

In its Alphabetical view, the list can be sorted on the tag or the word.
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).

![Colours screenshot]

Then in WordList, I chose View | Layout as in this screenshot, selected the Word column header and chose green below.
9.25 WordList display

Each WordList display shows

- the word
- its frequency
- its frequency as a percent of the running words in the text(s) the word list was made from
- the number of texts each word appeared in
- that number as a percentage of the whole corpus of texts

The Frequency display might look like this:
Here you see the top 7 words in a word list based on 480 texts. There are 72,028 words altogether but in the screenshot we can only see the first few. The Freq. column shows how often each word cropped up (THE looks as if it appeared 72,010 times in the 480 texts), and the % column tells us that the frequency represents 6.07% of the running words in those texts. The Texts column shows that THE comes in 480 texts, that is 100% of the texts used for the word list.

If we pull the Freq. column a little wider (cursor at the header edge then pull right) so that the 72,010 doesn't have any purple marks beside it,

we see the true frequency value is actually 172,010.

Another thing to note is that there seems to be a word #, with over 50 thousand occurrences.

That represents a number or any word with a number in it such as EX658.
The Alphabetical listing also shows us some of the words but now they're in alphabetical order. **ABANDON** comes 43 times altogether, and in 37 of the 480 texts (less than 8%). **ABANDONED**, on the other hand, not only comes more often (78 times) but also in more texts (14% of them).

Now let's examine the statistics.
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 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.

See also: Consistency, Lemmatisation

9.26 WordSmith controller: WordList settings

These are found in the main Controller marked WordList. This is because some of the choices -- e.g. Minimum & Maximum Settings -- may affect other Tools. There are 2 sets: *What you Get* and *What you See*. 
WHAT YOU GET

Word Length & Frequencies
See Minimum & Maximum Settings.

Standardised Type/Token #
See WordList Type/Token Information.

Detailed Consistency
Min. frequency overall = a total frequency for a word to be included in the Detailed Consistency list.
Min. texts = the minimum number of texts that word must appear in.

WHAT YOU SEE
By default you get “words only, no tags”. If you want to include tags in a word list, you need to set up a Tag File first. Then choose one of the options here.

In the example here we see that BECAUSE is classified by the BNC either as a `<w CJS>` or a `<w PRP>`. (That's how the BNC classifies BECAUSE OF...)

Tags

By default you get “words only, no tags”. If you want to include tags in a word list, you need to set up a Tag File first. Then choose one of the options here.

In the example here we see that BECAUSE is classified by the BNC either as a `<w CJS>` or a `<w PRP>`. (That's how the BNC classifies BECAUSE OF...)
Columns

The Columns to show/hide list offers all the standard columns: you may uncheck ones you normally do not wish to see. This will only affect newly computed KeyWords data: earlier data uses the column visibility, size, colours etc already saved. They can be altered using the Layout menu option at any time.

Case Sensitivity

Normally, you’ll make a case-insensitive word list. If you wish to make a word list which distinguishes between the, The and THE, activate case sensitivity.

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.
9.27 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 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 based on the index. A good maximum is probably 5 or 6.

*stop at:* you can choose where you want cluster breaks to be assumed. With the setting above (no limits), “I wrote the letter. Then I posted it” would consider letter
then I posted as a possible multi-word string even though there's a sentence break between them.

**Relationships**

See relationships computing.
10 Utility Programs

10.1 Convert Data from Previous Versions

10.1.1 Convert Data from Previous Versions

As WordSmith Tools develops, it has become necessary to store more data along with any given word-list, concordance etc. For example, data about which languages(s) were selected for a concordance, notes now stored with every type of results file, etc. Therefore it has been necessary to supply a tool to convert data from the formats used in WS 1.0 to 3.0 (last millennium) to the new format for the current version.

If you try to open a file made with a previous version you should be offered a chance to convert it first.

Note: as WordSmith develops, its saved data may get more complex in format. A concordance saved by WordSmith 5.0 cannot be guaranteed to be readable by WordSmith 4.0 for that reason, and a 6.0 one may require version 6.0, etc.

10.2 WebGetter

10.2.1 overview

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

What you do
Just type a word or phrase, check the language, and press Download.

How it works

WebGetter visits the search engine you specify and downloads the first 1000 sources or so. Basically it uses the search engine just as you do yourself, getting a list of useful references. Then it sends out a robot to visit each web address and download the web page in each case (not from the search engine’s cache but from the original web-site). Quite a few robots may be out there searching for you at once -- the advantage of this is that one slow download doesn't hold all the
After downloading a web page, that WebGetter robot checks it meets your requirements (in Settings) and cleans up the resulting text. If the page is big enough, a file with a name very similar to the web address will be saved to your hard disk.

When it runs out of references, WebGetter re-visits the search engine and gets some more.

See also: Settings, Display, Limitations.
10.2.2 settings

**Language**
Choose the language you require from the drop-down list.

**Search Engine**
The search engine box allows you to choose for example www.google.com.br for searches on Brazilian Portuguese or www.google.fr for French. That is a better guarantee of getting text in the language you require!

**Folder and Time-out**
where the texts are to be stored. By default it is the \swsmith5 folder stemming from your
My Documents. The folder you specify will act as a root. That is, if you specify c:\temp
and search for "besteirol", results will be stored in c:\temp\besteirol. If you do another
search on say "WordSmith Tools", results for that will go into c:\temp\WordSmith
Tools.

• timeout: the number of seconds after which WebGetter robot stops trying a given webpage if
there's no response. Suggested value: 50 seconds.

Requirements

• minimum file length (suggested 20Kbytes): the minimum size for each text file downloaded
from the web. Small ones may just contain links to a couple of pictures and nothing much
else.

• minimum words (suggested: 300): after each download, WebGetter goes through the
downloaded text file counting the number of words and won't save unless there are enough.

• required words: you may optionally type in some words which you require to be present in
each download; you can insist they all be present or any 1 of these.

Clean-up

If you want all the HTML markup removed, you can check this box, setting a suitable span between
< and > markers, 1000 recommended.

Advanced Options

If you work in an environment with a "Proxy Server", WebGetter will recognise this automatically
and use the proxy unless you uncheck the relevant box. If in doubt ask your network administrator.
You can specify the whole search URL and terms string yourself if you like with a box in the
Advanced options.

See also: Display  Limitations

10.2.3 display

As WebGetter works, it shows the URLs visited. If greyed out, they were too small to be of use or
haven't been contacted yet.

There is a tab giving access to a list of the successfully downloaded files which will show something
like this.
Double-click a file to view and, if you like, edit it in Notepad.

The URLs list looks like this
10.2.4 limitations

Everything depends on the search engine and the search terms you use. The Internet is a huge noticeboard; lots of stuff on it is merely ads and catalogue prices etc. The search terms are collected by the search engines by examining terms inserted by the web page author. There is no guarantee that the web pages are really "about" the term you specify, though they should be roughly related in some way.

Use the Settings to be demanding about what you download, e.g. in requiring certain words or phrases to be present.

See also: Display
10.3 Corpus Corruption Detector

10.3.1 Aim

The purpose is to check whether one or more of your text files in your corpus doesn't belong. This could be because:

- It has got corrupted so what used to be good text is now just random characters or has got cut much shorter because of disk problems.
- It isn't even in the same language as the rest of the corpus.

The tool works in any language. It does it by using a known sample of good text (in whatever language) and comparing that good text with all your corpus.

See also: How to do it

10.3.2 How it works

1. Choose a set of "known good text files" which you're sure of. The program uses these to evaluate the others.

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 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 Minimal Pairs

10.4.1 aim

A program for finding possible typos and pairs of words which are minimally different from each other (minimal pairs). For example, you may have a word list which contains ALREADY 5 and ALREADY 461, that is, your texts contain 5 instances where there is a possible misprint and 461 which are correct. This program helps to find possible variants and typos and anagrams.

See also: requirements, choosing your files, output, rules and settings, running the program.
10.4.2 requirements

A word-list in text format. Each line should contain a word and its frequency separated by tabs, e.g.

```
THE    224,637
AND    163,141
OF      119,980
```

or

```
1  THE    224,637
2  AND    163,141
3  OF      119,980
```

You can make such a list using WordList. For example, select (highlight) the columns containing the word and its frequency and

```
index file: Dickens

<table>
<thead>
<tr>
<th></th>
<th>Word</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>THE</td>
<td>224,637</td>
</tr>
<tr>
<td>2</td>
<td>AND</td>
<td>163,141</td>
</tr>
<tr>
<td>3</td>
<td>OF</td>
<td>119,980</td>
</tr>
<tr>
<td>4</td>
<td>TO</td>
<td>119,882</td>
</tr>
<tr>
<td>5</td>
<td>A</td>
<td>102,480</td>
</tr>
</tbody>
</table>
```

copy to the clipboard, then paste into Notepad, or

save as TXT (without numbers or heading row):
See also: aim, choosing your files, output, rules and settings, running the program.

10.4.3 choosing your files

Choose your input word list (which must be in plain text format) by clicking the button at the right of the edit space and finding the word list .txt file.
Type in an appropriate file-name for your results. Choose the rules also. 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, requirements, output, rules and settings, running the program.
10.4.4 output

An example of output is

418 ALTHOUGH (7) ALTHOUGH (37975)
Here the lines are numbered, and the bracketed numbers mean that ALTHOUGH 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, requirements, choosing your files, rules and settings, running the program.
10.4.5 rules and settings

![Minimal Pairs](image)

**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!

\textit{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.

\textit{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.

\textit{Factory Defaults} (restores default values)

See also: \textit{aim}, \textit{requirements}, \textit{choosing your files}, \textit{output}, \textit{running the program}.

\section*{10.4.6 running the program}

- Press "Compute".
  You should then see your source text, with a few lines visible. Some of the rows and columns may be greyed and others white: move the column and row numbers till the real data are white and any headings or line-numbers are greyed out.
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, requirements, choosing your files, output, rules and settings

10.5 File Viewer

10.5.1 Using File Viewer

Aim
To help you examine files of various kinds to see what is in them. This might be
- in order to see whether they’re really in plain text format
- to see whether there’s something wrong with them, such as unusual characters which oughtn’t to be there
- to see whether they were formatted for Windows, Mac, or for Unix
to check out any hidden stuff in there. (A *.doc* file 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.

### Loading a text file

Choose your file – if necessary click on the button at the right of the text-input box. Press Show.

### Characters

The two options available are as 1 bytes or 2 to represent each character-symbol in the text in question. You may need to alter this setting to see your text in a readable format.

### The two windows

The left window shows how the "text" is built up. You can see each character as a number and, further to the right, as a character.

The right window shows the text, paragraph by paragraph, word-wrapped so you can read it.

### Searching

Just type in the search-word and press Search. The search is case sensitive and is not a "whole word" search.

### Synch

Press the Synch button to synchronise the two displays. The display you clicked last is the "boss" for synchronising.

### Settings
Colours

The colour grids let you see the number section in special colours, so you can find the potential problems you're interested in.

- First select the character you want coloured.
- Click the foreground or background colour list change the colour.

The character names are Unicode names. In the picture the symbol with the 003E code (>) is the last one clicked.

Font

Choose the font and size in the font window. You may need to change font if you want to see Chinese etc. represented correctly.

Columns

- You can set the “hex” columns between 2 and 16.
- You can see the numbers at the left of the main window in hex or decimal.

10.6 File Utilities
10.6.1 index
This sub-program supplies a few file utilities for general use:

- Compare Two Files
- File Chunker
- Find Duplicates
- Rename
- Find Holes: for "holes" in text files
- Splitter
- Joiner
- Move files to sub-folder

10.6.2 Splitter

10.6.2.1 Splitter: index

Explanations

What is the Splitter utility and what's it for?

Filenames

Wildcards

See also: WordSmith Main Index

10.6.2.2 aim of Splitter

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

Start/End of Section Separator

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

Restrictions:
- 1. The start/end-of-text marker must occur at the beginning of a line in the original large file.
- 2. It is case sensitive: </Text> will not find </text>.
- 3. The first character in the separator may not be a wildcard such as #,* or ?.
- 4. * and # may occur only once each in the separator.

Splitter will create a new file every time it encounters the start/end-of-text marker you've specified. The end of text box determines whether the line containing the separator gets included in the previous or new text file.

Destination Folder

Where you want the small files to be copied to. (You'll need write permission to access it if on a network.)

Required sizes

The minimum and maximum number of lines that your small files can have (default = 5 and
30,000). Only files within these limits will be saved. This feature is useful for extracting files from very large CD-ROM files.
A "line" means from one <Enter> to the next.

**Bracket first line**

Whether or not you want the first line of each new text file to be bracketed inside < > marks. (If your separator is a start-of-section separator like CHAPTER with a number, you may wish that to be in brackets. And often the first line after an end-of-text symbol will contain some kind of header.) If you don't want it to insert < and > around the line, leave this box unchecked.

**Title Line**

If you know that a given line of your texts always contains the title for the sub-text in question, set this counter to that number, otherwise leave it at 0. For example, where you know that every line immediately following *<end of story>* has a title for the next story, you could put 1.
Example :

...<end of story>
Visiting New York
...

The file-name created for each story will contain the title as well as a suitable number. In this example a file-name might end up as *C:/texts/split/Visiting New York 0004.txt*.

See also: Joiner, Filenames, Wildcards, The buttons, Text Converter index.

**10.6.2.3 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 [38].
3. CD-ROM files when copied to your hard disk may be read-only. You can change this attribute using Text Converter [37].

**10.6.2.4 Splitter: wildcards**

- The hash symbol, #, is used as a wildcard to represent any number, so [FF#] would find [FF3] or [FF987] 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>].
10.6.3 join text files

This is a sub-program for joining small text files into bigger ones. You might want this because you aren’t interested in the different texts individually but are only interested in studying the patterns of a whole lot of texts.

When you choose Joiner you will see something like this:

**End of text marker**
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\guardianJoined.

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, Text Converter index.

10.6.4 compare two files

The point of it
The idea is to be able to check whether 2 files are similar or not. You may often make copies of files
and a few weeks later cannot remember what they were. Or you have used File Chunker to copy
a big file to floppies and want to be sure the copy is identical to the original.

This program checks whether
a) they are the same size
b) they have the same contents
(it goes through both, byte by byte, checking whether they match)
c) they have the same attributes
(file attributes can be "read only" [you cannot alter the file], "system" [a file which Windows
thinks is central to your operating system], "hidden" [one which is so important that Bill Gates
may be reluctant to even let you know it exists on your disk])
d) they have the same time & date.
How to do it
Specify your 2 files and simply press “Compare”.

See also: file chunker, find duplicates, rename

10.6.5 file chunker

The point of it
The idea is to be able to cut up a big file into pieces, so that you can copy it in chunks e.g. for emailing.
Naturally, you may later want to restore the chunks to one file.

How to do it: to copy a file
1. Specify your “file to chunk” (the big one you want to copy)
2. Specify your “drive & folder” (where you want to copy the chunks to)
3. Specify the “size of each chunk”
4. Specify whether to “compress while chunking” (compresses the file as it goes along)
5. Press “Copy”.

How to do it: to restore a file
1. Specify your “first chunk” (the first chunk you made using this program)
2. Specify which folder to “restore to” (where you want the results)
3. Specify whether to “delete chunks afterwards” (if they are not needed)
4. Press “Restore”.

See also: compare two files, find duplicates, rename

10.6.6 find duplicates

The point of it
The idea is to be able to check whether you have files with the same name in different folders. You may often make copies of files and a few weeks later cannot remember where they were.

By default this program only checks whether the files it is comparing have the same name but dates and file-size can be compared too. It handles lots of folders, the point being to locate unnecessarily duplicated files or confusing reuse of the same filenames.

How to do it
Specify your Folder 1 and simply press “Search”. Find Duplicates will go through that folder and any sub-folders and will report any duplicates found.
Or you can specify 2 different folders (e.g. on different drives) and the process compares one set with the other.

**Sub-folders to exclude**

Useful if there are some sub-folders you know you're not interested in. In the example below, any folder whose name ends _old or _shibs or whose name is demo or examples will be ignored as will any sub-folder below it.

In the window below, you will find all the duplicates listed with the folder and date. In the example we can see there are two files called ambassador 1.txt in different shakespeare folders.

See also: compare two files, file chunker, rename

### 10.6.7 rename

**The point of it**

To rename a lot of files at once, in one or more folders. You may have files with excessively long names which do not suit certain applications. Or it is a pain to rename a lot of files one by one.
The idea is to rename a set of files with a standard name plus a number. For example suppose you have downloaded a lot of text files containing emails from the Enron scandal, you could rename them

```
Enron001.txt
Enron002.txt
etc.
```

**How to do it**

![Image of the interface](image)

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

In the screenshot, *.txt;*.xml has been specified, which means all .txt files and all .xml files. Find Files has been pressed, too. In the list you can see some of each.

If you typed baby?? .doll you'd get all files with the .doll ending as long as the first 4 characters were baby as in baby05.doll, babyyz.doll, etc.

Now specify a "mask for new name" and a starting number. The mask can end with a series of # characters standing for numbers. In this screenshot, there are 4 # symbols.
so after pressing Rename the texts have been renamed Bacon plus an incrementing number formatted to 4 digits.

See also: compare two files, file chunker, find duplicates

10.6.8 move files to sub-folders

This function allows you to take a whole set of files in a folder and move them to suitable sub-folders.

Example:

In c:\temp you have

2001 Jan.txt
2001 Feb.txt
2003 Jan.txt
2003 Feb.txt
2003 March.txt
2003 Oct.txt

e tc. and you want them sorted by year into different folders.

Using the template AAAA* you will take the first four characters of your files and place each into a sub-folder named appropriately.

Results

c:\temp\2001 contains 2001 Jan.txt, 2001 Feb.txt
and all the others are in c:\temp\2003

Syntax
? = ignore this character
A = use this character in the file-name
* = use no further characters in the file-name

10.6.9 dates and times

Purpose
The aim here is to parse your file-names identifying suitable textual dates and times where you have incorporated something suitable in the file-name. Suitable dates can be re-used by saving file-choices as favourites.

Mask Syntax
The procedure reads any file-names in the Folder to process (and optionally its sub-folders) and attempts to parse them. If an indicator is found it will record a suitable date combination. Suitable indicators of textual date are

YY or YYYY year, two or four digits (YY=a 20th Century date)
MM month
DD day
* skip all characters until a digit is found

The procedure doesn't understand words such as "December" or "Five", it only uses digits. Any character other than Y,M,D,* in the mask simply gets ignored.

Output
The program will always add each entry found to a simple text file (File for list of dates) listing its file-name and adding a suitable date as expected in the auto-date procedure[47], (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[47] within file-choosing and save your preferred text files as favourites.

Examples

<table>
<thead>
<tr>
<th>Your Mask</th>
<th>Source file</th>
<th>Date and Time interpreted</th>
</tr>
</thead>
<tbody>
<tr>
<td>YYYYMMDD</td>
<td>20060512 Peter monologue.txt</td>
<td>12th May 2006 (first 8 characters used in the mask)</td>
</tr>
<tr>
<td>YYMMDD</td>
<td>841231.txt</td>
<td>31st December 1984 (20th Century assumed if YY mask used)</td>
</tr>
<tr>
<td>DDMMYY</td>
<td>311284.txt</td>
<td>31st December 1984</td>
</tr>
<tr>
<td>DDMYYYY</td>
<td>20060512 Peter monologue.txt</td>
<td>20th June, the year 512 AD</td>
</tr>
<tr>
<td>DDMY</td>
<td>20060512 Peter monologue.txt</td>
<td>20th June of the current year</td>
</tr>
<tr>
<td>#YYYYMMMPeter</td>
<td>20060512.txt</td>
<td>12th May 2006 (first six characters were ignored, five for Peter, one for space)</td>
</tr>
<tr>
<td>DD</td>
<td>*YYYY Peter 20060512.txt</td>
<td>15 July 2006 (all characters to first digit skipped, then next 4 used for year date)</td>
</tr>
<tr>
<td>YYYY</td>
<td>1086 Domesday book.txt</td>
<td>15 July 1086 (there were only four digits)</td>
</tr>
<tr>
<td>YYYYMMMD</td>
<td>1086 Domesday book.txt</td>
<td>15 July 1086 (mask had 8 digits but file-name only 4)</td>
</tr>
<tr>
<td>YYYY#MM#DD</td>
<td>2006,05/12,10-54.txt</td>
<td>12th May 2006</td>
</tr>
<tr>
<td>YYYY MM DD</td>
<td>2006,05/12,10-54.txt</td>
<td>12th May 2006</td>
</tr>
</tbody>
</table>
10.6.10 find holes in texts

After text files have been copied from one source to another, they may get slightly corrupted with holes in the stream of text. This utility lets you seek out the texts in your corpus which have got corrupted in this way and optionally lets you delete them. If you want to convert the holes to space-characters, use the Text Converter.

10.7 Text Converter

10.7.1 purpose

This program does a "Search & Replace", on virtually any number of files.

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

converting text

For a simple search-and-replace you can type in the search item and a replacement; for more complex conversions, use a Conversion File so that Text Converter knows which symbols or strings to convert. It operates under Windows and saves using the Windows character set, but will convert text using DOS or Windows character sets. You can use it to make your text files suitable for use with 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 specified, the Text Converter will read each source file and either create a new version or replace the old one, depending on the overwrite setting.

You will be able to see the details of how many instances of each string were found and replaced overall.

filtering files

And/or you may need to make sure texts which meet certain criteria are put into the right folders.

Tip

The easiest way to ensure your text files are the way you want, especially if you have a very large number to convert, is to copy a few into a temporary folder and try out your conversion file with the Text Converter. You may find you've failed to specify some necessary conversions. Once you're sure everything is the way you want it, delete the temporary files.
10.7.2 index

Explanations
- What is the Text Converter and what's it for?
- Getting Started...
- Convert the text format
- Filters
- Sample Conversion File
- Syntax
- Conversion File

See also: WordSmith Main Index

10.7.3 settings

1. Choose Files (the top left tab). Decide whether you want the program to process sub-folders of the one you choose. There is no limit to the number of files Text Converter can process in one operation.
2. Click on the Conversion or Filters 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 responsible becomes responsible) or you can choose your own pre-prepared Conversion File.

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 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.
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 \[\text{Filters}\] 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 \[\text{copy to}\] 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.

**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.

**10.7.3.1 Text Converter: copy to**

If you choose to copy the files you are converting, instead of converting or filtering them in place, which is a lot safer, the new files created will be structured like this.

Suppose you are processing `d:\texts\2007\literature` and copying to `c:\temp` and suppose `d:\texts\2007\literature` contains this sort of thing:

```
  d:\texts\2007\literature\shakespeare\hamlet.pdf
  d:\texts\2007\literature\shakespeare\macbeth.pdf
  ...
  d:\texts\2007\literature\shakespeare\poetry\sonnet1.pdf
  d:\texts\2007\literature\shakespeare\poetry\sonnet2.pdf
  ...
  d:\texts\2007\literature\french\victor hugo\miserables.pdf
  d:\texts\2007\literature\french\poetry\baudelaire\le chat.pdf
  ...
```

you will get

```
  c:\temp\shakespeare\hamlet.txt
  c:\temp\shakespeare\macbeth.txt
  ...
```
In other words, for each file successfully converted or filtered, any same directory structure beyond the starting point (d:\texts\2007\literature in the example above) will get appended to the destination.

### 10.7.4 extracting from files

**The point of it...**

The idea is to be able to extract something useful from within larger files. In the example below, I wanted to extract the headlines only from some newspaper text. I knew that the header for each text contained `<DAT>` (date of publication mark-up) and that the headline ended `</HED>`, and I wanted only those chunks which contained the phrase *leading article*:

The results I got looked like this:

```xml
<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>
```
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.

10.7.5 filtering: move if

This function allows you to specify a word or phrase, look for it in each file, and if it's found move that file into a new folder.

The point of it ...

Suppose you have a whole set of files some of which contain dialogues between Pip and Magwitch, others containing references to the Great Wall of China or the anatomy of fleas. You want those with the Pip-Magwitch 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 World Edition there are no file extensions) and/or convert it to Unicode.
- You could also have any texts not containing the word *Magwich* copied to a specified folder.

The *load BNC World* and *load BNC XML* buttons are specific to those two editions of the BNC and
read text files with similar names which you will find in your Documents\wsmith6 folder.

See also: Text Converter Contents.

10.7.6 Convert within the text file

Your choices here are 5:
1. cut out a header

and/or
2. make one change only
3. insert numbering
4. replace some problem characters
5. use a script to determine a whole set of changes. There is an sample to see.

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 \(<\text{Enter}>\) (line-break or paragraph end), to treat the search as literal (see \textit{syntax}) and to force the results into Unicode.

It might be best to check the \textit{confirm each} box too if there's any danger of confusing two different Dorothies with each other.

### Insert numbering

This allows you to insert paragraph numbering into your corpus texts. When you click the \textit{specify numbering} button you'll get options like these:

![Insert Numbering dialog](image)

With these choices, for each of your texts, a string like \textit{<para "1">}, \textit{<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 \textit{Ulan Bator}.

Paragraphs here are identified simply as sequences ending in one \(<\text{Enter}>\).

### 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 \textit{with this character} space blank.

See also: \textit{convert whole file}, \textit{sample conversion file}, \textit{syntax}, \textit{Text Converter Contents}.

#### 10.7.6.1 conversion file

Prepare your Text Converter conversion file using a plain text editor such as Notepad. You could use \texttt{Documents\wsmith6\convert.txt} as a basis.

If you have \textit{accented characters} in your original files, use the DOS editor to prepare the conversion file if they were originally written under DOS and a Windows editor if they were written in a Windows word-processor. Some Windows word processors can handle either format.

There can be any number of lines for conversion, and each one can contain two strings, delimited
with " " quotes, each of up to 80 characters in length.
The Text Converter makes all changes in order, as specified in the Conversion File. Remember one alteration may well affect subsequent ones.

**Alterations that increase the original file**
Most changes reduce the size of an original. But Text Converter will cope even if you need to increase the original file -- as long as there's disk space!

**Tip**
To get rid of the <Enter> at line ends but not at paragraph ends, first examine your paragraph ends to see what is unique about them. If for example, paragraphs end with two <Enter>s, use the following lines in your conversion file:

```
"{CHR(13)}{CHR(10)}{CHR(13)}{CHR(10)}" -> "{%%}"
```

(this line replaces the two <Enter>s 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 <Enter>s with a space, to keep words separate)

```
"{%%}" -> "{CHR(13)}{CHR(10)}{CHR(13)}{CHR(10)}"
```

(this line replaces the {%%} combination with <Enter><Enter>, thus restoring the original paragraph structure)

```
/S
```

(this line cuts out all redundant spaces)

See also: sample conversion file, syntax, Text Converter Contents.

10.7.6.2 syntax

The syntax for a Conversion File is:

- Only lines beginning / or " are used. Others are ignored completely.
- Every string for conversion is of the form "A" -> "B". That is, the original string, the one you're searching for, enclosed in double quotes, is followed by a space, a hyphen, the > symbol, and the replacement string.
- 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 for more.
Wildcards

The search uses the same mechanism that Concord uses. You may use the same wildcards as in Concord search-word syntax. 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 on either side or not) (/U "the" finds the but also finds other, then and bathe).

/I does a case insensitive search (/I "restaurant" -> "hotel" replaces restaurant with hotel and RESTAURANT with HOTEL and Restaurant with Hotel, i.e. respecting case as far as possible).

You can combine these, e.g.

/IC "the" -> "this"

/S cuts out all redundant spaces. That is, it will reduce any sequence of two or more spaces to one, and it also removes some common formatting problems such as a lone space after a carriage-return or before punctuation marks such as .; and ). /S can be used on a line of its own or in combination with other searches.

/E cuts out all redundant <Enter>s. That is, it will reduce any sequence of two or more carriage-return+line-feeds (what you get when you press Enter or Return) to one. /E can be used on a line of its own or in combination with other searches.

/L means both the search and replace strings are to be taken as literal. (Normally a sequence like <#*> would need quotes around each character because < > are mark-up signals and # and * are special wildcard characters, thus "<""#"""">" which is tricky! Put /L at the start of the line to avoid this.)

See Documents\wsmith6\convert.txt to see examples in use.

See also: Text Converter Contents [347].

10.7.6.3 sample conversion file

You could copy all or part of this to the clipboard and paste it into notepad.

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

[ first a spelling correction ]

"responsable" -> "responsible"
10.7.7 Convert format of entire text files

To convert a series of whole text files from one format to another, choose one or more of these options:
These formats allow you to convert into formats which will be suited to text processing.

- **into Unicode:**

  .... this is a better standard than ASCII or ANSI as it allows many more characters to be used, suiting lots of languages. See Flavours of Unicode.
TXT file extensions:

... makes the filename end in .txt (so that Notepad will open without hassling you; Windows was baffled by the empty filenames of the BNC editions prior to the XML edition). If you choose this you will be asked whether to force .txt onto all files regardless, or only ones which have no file extension at all.

curly quotes etc.:

... changes any curly single or double quote marks or apostrophes into straight ones, ellipses into three dots, and dashes into hyphens. (Microsoft's curly apostrophes differ from straight ones.)

removing line-breaks

... replaces every end of line line-break with a space. Preserves any true paragraph breaks, which you must ensure are defined (default = <Enter><Enter> -- in other words two line-breaks one after the other with no words between them).

See also: Mark-up, Word/Excel/PDF, non-Unicode, convert within text files, MS Word documents, Guide to handling the BNC
10.7.1 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 for looking for the next > when it finds a <.

**word_TAG to <TAG>word**

The Helsinki corpus can come tagged like this (COCOA tags)

```
the_D occasion_N of_P her_PRO$ father's_N$ death_N
```

and this conversion procedure will change it to

```
<D>the <N>occasion <P>of <PRO$>her <N$>father's <N>death
```

Note: this procedure does not affect underscores within existing <> markup.

**word_TAG to word<TAG>**

cconverts text like
It_PP is_VBZ easy_JJ

or Stanford Log-linear POS tagger output like
It/PP is/VBZ easy/JJ

to
It<PP> is<VBZ> easy<JJ>

You will have to confirm which character such as _ or / divides the word from the tags. Note: before
it starts, it will clear out any existing <> markup.

**swap tag and word**

converts text like
It<PP> is<VBZ> easy<JJ>

to
PP><It VBZ>is <JJ>easy

or vice-versa. In other words swapping the order of tags and words. The procedure effects a swap at
each space in the non-tagged text sequence.

Any tags which do not qualify a neighbouring word but for example a whole sentence or a paragraph
should not be swapped, so fill in the box to the right with any such tags, using commas to separate
them, e.g. <s>,</s>,<p>,</p>

**from column tagged**
The Stuttgart Tree Tagger produces output like this separating 3 aspects of each word with a <tab>:

<table>
<thead>
<tr>
<th>word</th>
<th>pos</th>
<th>lemma</th>
</tr>
</thead>
<tbody>
<tr>
<td>The</td>
<td>DT</td>
<td>the</td>
</tr>
<tr>
<td>TreeTagger</td>
<td>NP</td>
<td>TreeTagger</td>
</tr>
<tr>
<td>is</td>
<td>VBZ</td>
<td>be</td>
</tr>
<tr>
<td>easy</td>
<td>JJ</td>
<td>easy</td>
</tr>
<tr>
<td>to</td>
<td>TO</td>
<td>to</td>
</tr>
<tr>
<td>use</td>
<td>VB</td>
<td>use</td>
</tr>
<tr>
<td>.</td>
<td>SENT</td>
<td>.</td>
</tr>
</tbody>
</table>

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\text{<Treetagger><NP> is<be><VBZ> etc.}

- the template \text{<POS="2">/1} will produce \text{<POS="DT">The <POS="NP">Treetagger <POS="VBZ">is etc.}

It will present the text as running text, no longer in columns, but with a break every 80 characters.

**entities to characters**

... converts HTML or XML symbols which are hard to read such as \text{sacute;} to ones like \text{é.}

Specify these in a text file. There is a sample file pre-prepared for you, \text{html_entities.txt}, in your Documents\wsmith6 folder; look inside and you'll see the syntax.

**XML simplification**

The idea is to remove any mark-up in XML data which you really do not wish to keep. For example, in the BNC XML edition you might wish to keep only the \text{pos="*"} mark-up and remove the \text{c5} and \text{hw} attributes.

```xml
<w c5="PRP" hw="in" pos="PREP">In</w>
<w c5="ART" hw="the" pos="ART">the</w>
<w c5="ADJ" 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="VVZ" hw="have" pos="VERB">has</w>
<w c5="VNP" 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="ART" hw="the" pos="ART">the</w>
<w c5="AJJ" hw="big" pos="ADJ">biggest</w>
<w c5="NN1" hw="gold" pos="SUBST">gold</w>
<w c5="VVZ" hw="rush" pos="VERB">rushes</w>
<w c5="BNP" hw="it" pos="PRON">it</w>
```

To do so, press the Options button and complete for example like this:

![Simplify XML](image)

resulting in a saved XML file with a structure like this:
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

10.7.7.2 Word, Excel, PDF

from MS Word or Excel to .txt
This is like using "Save as Text" in Word or Excel. Handles .doc, .docx (Office 2007) and .xls files.

### from PDF

... into plain text. Not guaranteed to work with every .PDF as formats have changed and some are complex.

To convert PDFs to plain text can be extremely tricky even if you own a licensed copy of the Adobe software (Adobe themselves created the PDF format in 1993). That is because PDF is a representation of all the dots, colours, shapes and lines in a document, not a long string of words. It can be very hard with an image of the text, to determine the underlying words and sentences. A second problem is that PDFs can be set with security rights preventing any copying, printing, editing etc. Other formats (.TXT, .DOC, .DOCX, .XML, .HTML, .RTF etc.) are OK in principle as they do not contain only an image but also store within themselves the words and sentences.

See also: Convert Entire Texts

#### 10.7.7.3 non-Unicode Text

![non-Unicode Text](image)
**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](#)

### 10.7.7.4 Other changes

#### Unix to Windows
Unix-saved texts don't use the same codes for end-of-paragraph as Windows-saved ones.

**encrypting using**
... allows you to encrypt your text files. You supply your own password in the box to the right. When WordSmith processes your text files, e.g. when running a concordance it will restore the text as needed but otherwise the text will be unintelligible. Encrypted files get the file extension `.WSencrypted`. For example, if your original is `wonderful.txt` the copy will be `wonderful.WSencrypted`. Requires the safer `copy to` button above to be selected.
lemmatising using

... converts each file using a lemma file. 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?"

See also: Convert Entire Texts

10.7.8 Text Converter: converting BNC XML version

The British National Corpus is a valuable resource but has certain problems as it comes straight off the cdrom:

- it is in Unix format
- it has entities like 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 http://lexically.net/wordsmith/Handling_BNC/index.html

The easiest way to do that is in two stages.

Conversion:
After choosing the texts,
and when you press OK you'll be asked something like this

![Confirm dialog box]

4,049 texts to process including 4,049 in sub-folders
-- Filters: none
Conversion: from Unix, into Unicode, from HTML entities, adding .TXT?

[Options: Cancel, Yes to All]
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:

![Filtering Interface](image)

Eventually you should get folder structures like this:
10.8 Viewer and Aligner

10.8.1 purpose

This is a program for showing your text or other files, highlighting words of interest. You will see them in plain text format, with tag mark-up shown or hidden as in your tag settings. There are a number of settings and options you can change.

Its main use is to produce an aligned version of 2 or more texts, with alternate sentences or paragraphs from each of them.
What is the Viewer & Aligner and what’s it for? [373]

an example of aligning [379]

Viewing Options [379]

What to do if it doesn’t do what I want… [428]

Searching for Short Sentences [380]

Joining/Splitting [380]

Aligning a Dual Text [377]

Finding translation mis-matches [380]

The technical side… [380]

See also: Viewer & Aligner settings, Viewer & Aligner options, an example of aligning
10.8.3 aligning with Viewer & Aligner

This feature aligns the sentences in two files. Translators need to study differences between an original and a translation. Other linguists might want it to study differences between two versions of a text in the same language. Students of different languages 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, Aligning and moving

10.8.4 example of aligning

How to do it -- a Portuguese and English example

1. Read in your Portuguese text (eg. Hora da Estrela.TXT), and checking its sentences and paragraphs break the way you like. Try “Unusual Lines” to help identify oddities.
2. Save it

New... Ctrl+N
Open... Ctrl+O
Merge with... Ctrl+W
CloseCtrl+W
Save Ctrl+F2
Save As
Print... Ctrl+P
Print Preview F3
Exit Alt+F4
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*
7. Finally File | Save choosing Aligned files (.ALI) as the format.
10.8.5 aligning and moving

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

adjusting by dragging with the mouse

To merge sentences or paragraphs, simply grab and drag it up to the next one above in the same language. Or use the Join button. Or press F4.
To split a sentence or paragraph, choose the Split button or press Ctrl+F4.

Finally you will want to save (Ctrl+F2) the results.

See also: Viewer & Aligner contents

10.8.6 editing

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

- Split: allows you to choose where a line should be divided in two.
- Join down, Join up: these buttons merge a line with another one. You can achieve this also by simply dragging.
- Cut line: removes any blank lines.
- Trim: this goes through each sentence of the text, removing any redundant spaces -- where there
are two or more consecutive spaces they will be reduced to one.

- Cut & Trim All does these actions on the whole text.
- Edit opens up a window allowing you to edit the whole of the current sentence or paragraph.
- Heading: allows you to treat a line as a heading, and if so makes it look bold.
- Find unusual lines: this identifies cases where a sentence or paragraph does not start with a capital letter or number – you will probably want to join it to the one above, or where a line is unusually short, etc.
- Find short lines

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 (select, then press Control+Ins)
- print the whole or part of the currently active text file
- search for words or phrases (press F12)

10.8.7 languages

Each Viewer file (.VWR) has its own language. Each Aligner file (.ALI) has one language for each of the component sections. (They could all be the same, if for example you were analysing various different editions of a Shakespeare play they'd all be English.) The set of languages available is that defined using the Languages Chooser.

You can change the language to one of your previously defined languages using the drop-down list. Here is an example where a Portuguese language plain TXT text file was opened and the default language was English.

When Portuguese was chosen in the drop-down list, and
agreed to, it was possible to save the result (as a .VWR file) so that henceforth it would know which language to use.

10.8.8 numbering sentences & paragraphs

You can use the Viewer & Aligner to make a copy of your text with all the sentences and/or paragraphs tagged with <S> and <P>. To do this, simply read in the text file in, choose Edit | Insert Tags, then save it as a text file.

See also: Viewer & Aligner contents

10.8.9 options

**Mode: Sentence/Paragraph**

This switches between Sentence mode and Paragraph mode. In other words you can choose to view your text files with each row of the display taking up a sentence or a paragraph.

Likewise, you can make an dual aligned text by interspersing either paragraphs or sentences. The other functions (e.g. joining, splitting) work in the same way in either mode.

**Colours**

The various texts in your aligned text will have different colours associated with them. Colours can be changed using the button.
10.8.10 reading in a plain text

In Viewer and Aligner, choose File | Open, and select your plain text file.

and you may see this sort of thing in Sentence view,

or in Paragraph view,
Edit it, as necessary, e.g. **splitting or merging** 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:
10.8.11 joining and splitting

**Joining**

The easiest way to join two sentences is simply to drag the one you want to move onto its neighbour above. Or select the lower of the two and press F4 or use the button (_pose)

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 (Uncheck). 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*
10.8.12 settings

1. What constitutes a "short" sentence or paragraph (default: less than 25 characters)
2. Whether you want to do a lower-case check when Finding Unusual Lines

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

- Colours
- Font
- Printing
- Text Characteristics
- Review all Settings

10.8.13 technical aspects

**When is a sentence not a sentence?**

There is no perfect mechanical way of determining sentence-breaks. For example, a heading may well have no final full stop but would normally not be considered part of the sentence which follows it. And a sentence may often have no final full stop, if what follows it is a list of items.

The algorithm used by Viewer & Aligner is: a sentence ends when it meets the requirements explained in the definition of a sentence. The same routine is used as in WordList.

Consider this chunk from *A Tale of Two Cities*:

"Wo-ho!" said the coachman. "So, then! One more pull and you're at the top and be damned to you, for I have had trouble enough to get you to it! - Joe!"

**Viewer & Aligner** will mistakenly consider - Joe! as a separate sentence, but handles "Wo-ho!" said the coachman. as one: though the program would split it in two if the word after ho! had a capital letter (e.g. in *Wild Bill, the coachman, said.*)

**Viewer & Aligner** cannot therefore be expected to handle all sentence boundaries exactly as you would. (*I saw Mr. Smith.* would be considered two sentences; several headings may be bundled together as one sentence.) For this reason you can choose Find Short Sentences to seek out any odd one-word sentences.

See also: [Viewer & Aligner contents](#)

10.8.14 translation mis-matches

**Viewer & Aligner** can help find cases where alignment has slipped (one sentence having been translated as two or three). One method is to use the menu item *Match by Capitals.* This searches for matching proper nouns in the two versions: if say *Paris* is mentioned in sentences 25 of the source text and not in sentence 25 of the translation but in sentence 27, it is very likely that some slippage has occurred.

**Viewer & Aligner** will search forwards from the current text sentence on, and will tell you where
there's a mis-match. You should then search back from that point to find where the sentences start to diverge. It may be useful to sample every 10 or every 20 to speed up the search for slippage. When you find the problem, un-join or join and/or edit the text as appropriate, then save it.

See also: The technical side... Finding unusual sentences Viewer & Aligner contents

10.8.15 troubleshooting

**Can't see the whole sentence or paragraph**
Press to "auto-size" the lines in your display. This adjusts line heights according to the current highlighted column of data.

**Can't see the whole text file**
Press to "refresh" the display.

**Don't like the colours**
Change colours using . The colours initially used for each language version in the dual-language window are the same colours as used for primary sorting and secondary sorting in Concord.

See also: Viewer & Aligner contents

10.8.16 unusual lines

It can be useful to seek unusually short sentences to see whether your originals have been handled as you want. Because Viewer & Aligner uses full stops, question marks and exclamation marks as sentence-boundary indicators, you will find a string like "Hello! Paul! Come here!" is broken into 3 very short sentences. Depending on your purposes you may wish to consider these as one sentence, e.g. if a translator has translated them as one ("Oi, Paulo, venha cá!").

This function can also find lower-case lines: where a sentence or paragraph does not start with a capital letter or number -- you will probably want to join it to the one above. This problem is common if the text has been saved as "text only with line breaks" (where an <Enter> comes at the end of each line whether or not it is the end of a paragraph.)

**Seeking**
Use the Find Unusual Toolbar menu item and then press Start Search. Viewer & Aligner will go to the next possibly problematic sentence or paragraph and you will probably want to join 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, The technical side..., Finding translation mis-matches, Viewer & Aligner contents

10.9 WSCongram

10.9.1 aims

A program for finding concgrams, essentially related pairs, triplets, quadruplets (etc.) of words which are related.
10.9.2 definition of a concgram

For years it has been easy to search for or identify consecutive clusters (n-grams) such as AT THE END OF, MERRY CHRISTMAS or TERM TIME. It has also been possible to find non-consecutive linkages such as STRONG within the horizons of TEA by adapting searches to find context words. 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.
10.9.3 settings

The settings are found in the main Controller.

10.9.4 generating concgrams

To start, as usual, choose File | New.

In the Getting Started window, first choose an existing Index, as here where an index based on the works of Dickens has been selected.
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 a few million words, though, it should be possible to generate the files in a matter of a few minutes.

As you see above, some large additional files have been generated at the end of the two Build steps marked on the buttons in the top window.

All items which are found together at least as often as set in the Index settings (here 5 times)

will be saved as potential members of each concgram.

Now, choose Show to view the results. (Or, as usual, right-click the main WSConcgram window and choose last file).
10.9.5 viewing concgrams

When you first open a concgram file created by WSConcGram, it will look something like this one:

It'll appear (by default) in frequency order as set in the settings 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

![Tree View](image)

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 ( ), for filtering ( ), clearing filters ( ), and showing more or less of the tree ( ).

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 appropriately and pressing the Concordance button
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.

10.9.6 filtering concgrams

In order to select which items are "associated", we need some sort of suitable statistical procedures. The members of each concgram are at present merely associated by co-occurring at least a certain number of times as explained in generating them.

The Filtering settings in the Controller allow you to specify, for example, that you want to see only those which are associated with a MI (mutual information) score of 2.0 or a Log Likelihood score of 3.0.
Ensure the statistics you need are checked and set to suitable thresholds, and decide whether all the thresholds have to be met (in the case above both MI and Log Likelihood would have to score 3.0 at least) or any of them (in the case above MI at 3.0 or above or Log Likelihood at 3.0 or above). You can also optionally insist on certain words being in your filtered results.

When you press the filter button (✓), 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. To the right you see that the head-word CAESAR here relates to AND HE, HER, I, ANTONY etc. above the thresholds set.

10.9.7 exporting concgrams

With concgram data loaded, you may wish to export it to a plain text file which can be imported into Excel or imported into a WordSmith word-list.

Choose Compute | WordList and you will be offered choices like these.

The suggested filename is based on your concgram data.
10.10 Character Profiler

10.10.1 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:

<table>
<thead>
<tr>
<th>Top 10 characters</th>
<th>Bulgarian</th>
<th>Czech</th>
<th>English</th>
<th>Estonian</th>
<th>Hungarian</th>
<th>Latvian</th>
<th>Lithuanian</th>
<th>Romanian</th>
<th>Russian</th>
<th>Serbo-Croatian</th>
<th>Slovene</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a o e a e a i e o a e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>e e t e t i a a e o a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>o a a t a s s t a e t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>m n o s l e e r h i o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>h l i t n t t n m n n</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>t t t s n n n u u t s l</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>c s h u k u o t a r r</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>r v s k o r n c c t s</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>b i r n i k r s p j j</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>k k d d r l k l n u t</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

For further details see http://lexically.net/downloads/corpus_linguistics/1984_characters.xls.

10.10.2 profiling characters

How to do it

1. Choose one or more texts or a folder. You can type in a complete filename (including drive and folder), and can use wildcards such as *.txt, or you can browse to find your text or folder.
2. If you want to study one text only, just choose one text, but you may choose a whole folderful or more by using the "sub-folders too" option.
3. Press Analyse.
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**

<table>
<thead>
<tr>
<th>code</th>
<th>character</th>
<th>type</th>
<th>%</th>
<th>freq.</th>
<th>1st Position</th>
<th>2nd</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>&lt;Tab&gt;</td>
<td>Control</td>
<td>0.18 %</td>
<td>22,888</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>&lt;line-f... &gt;</td>
<td>Control</td>
<td>0.60 %</td>
<td>73,754</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>&lt;carria... &gt;</td>
<td>Control</td>
<td>0.60 %</td>
<td>73,754</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>space</td>
<td>SpaceS</td>
<td>17.05 %</td>
<td>2,109,771</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>!</td>
<td>punctua...</td>
<td>0.00 %</td>
<td>505</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>&quot;</td>
<td>punctua...</td>
<td>0.15 %</td>
<td>19,152</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>#</td>
<td>punctua...</td>
<td>0.00 %</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>$</td>
<td>Currency</td>
<td>0.00 %</td>
<td>427</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>%</td>
<td>punctua...</td>
<td>0.00 %</td>
<td>142</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>punctua...</td>
<td>0.01 %</td>
<td>1,172</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>punctua...</td>
<td>0.41 %</td>
<td>50,709</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>(</td>
<td>punctua...</td>
<td>0.04 %</td>
<td>5,070</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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,

Copy
Copies the data to the clipboard, ready to be pasted for example into Excel.

See also: settings

10.10.3 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 \textless\textit{Header}\textgreater, the program treats all the text before that as a header to be excluded from analysis..

Copy letter characters only
Check this one to force the copying to the clipboard to copy only data of letters, ignoring punctuation and digits.

Merge lower and UPPER case
Check this one to convert all text to upper case.
10.11 Chargrams

10.11.1 purpose

The point of it...

Chargrams, a tool to help find out which chargrams (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.

These are 3-letter chargrams occurring in word-final position. ING is a well-known ending in English; HAT is a frequent 3-letter sequence at the end of words too.

How does it work?

Chargrams are computed by taking only the valid characters of text. If a text contained "In 1845 there was a princess", the 3-character chargrams considered would be THE, HER, ERE, WAS, PRI, RIN, INC, NCE, CES, ESS. The positions are computed in relation to the original words, so THE is word-initial while ESS is word-final, and RIN is medial.

If including punctuation, the sequences would include IN_, N_1, _18, 184 etc. too.

10.11.2 chargram procedure

How to do it

Choose File | New in the Chargrams menu.

Choose your texts as with the other Tools.
Then press the button to make a chargram list.

See also: settings

10.11.3 display

The display is similar to that in the WordList tool.

Contexts

This column shows the word-contexts for each chargram. You can double-click to see the whole list.
### 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):

<table>
<thead>
<tr>
<th>context words</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>COU</td>
<td>15</td>
</tr>
<tr>
<td>COUCH</td>
<td>10</td>
</tr>
<tr>
<td>COUCHED</td>
<td>6</td>
</tr>
<tr>
<td>COUGH</td>
<td>7</td>
</tr>
<tr>
<td>COUGHED</td>
<td>6</td>
</tr>
<tr>
<td>COUGHING</td>
<td>5</td>
</tr>
<tr>
<td>COUGHS</td>
<td>4</td>
</tr>
<tr>
<td>COULD</td>
<td>4,068</td>
</tr>
<tr>
<td>COULDN'T</td>
<td>444</td>
</tr>
<tr>
<td>COULDVIE</td>
<td>24</td>
</tr>
<tr>
<td>COULSON</td>
<td>17</td>
</tr>
</tbody>
</table>
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

10.11.4 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, digits etc.) to be included. Spaces get replaced by underscores.

**Ignore low-frequency context chargrams**

This setting allows us to filter out any chargrams which do not occur in many contexts. As shown here, any chargrams not occurring in at least 4 context word types will get eliminated. If, for example, a chargram has been found in 5 context word-types then the chargram is included in your list. But if in only 1 of these it is found occurring at least 4 times (i.e. found in the same context word-type recurring in the texts at least 4 times), you will see one context word only in the contexts column.

See also: chargrams display [402]
Reference

Section XI
11 Reference

11.1 acknowledgements

WordSmith Tools has developed over a period of years. Originally each tool came about because I wanted a tool for a particular job in my work as an Applied Linguist. Early versions were written for DOS, then Windows came onto the scene.

One tool, Concord, had a slightly different history. It developed out of MicroConcord which Tim Johns and I wrote for DOS and which Oxford University Press published in 1993.

The first published version was written in Borland Pascal with the time-critical sections in Assembler. Subsequently the programs were converted to Delphi 16-bit; this is a 32-bit only version written in Delphi XE and still using time-critical sections in Assembler.

I am grateful to
• lots of users who have made suggestions and given bug reports, for their feedback on aspects of the suite (including bugs), and suggestions as to features it should have.
• generations of students and colleagues at the School of English, University of Liverpool, the MA Programme in Applied Linguistics at the Catholic University of São Paulo, colleagues and students at Aston University.
• Audrey Spina, Élodie Guthmann and Julia Hotter for their help with the French & German versions of WS 4.0; Spela Vintar's student for Slovenian; Zhu Yi and others at SFLEP in Shanghai for Mandarin for WS 5.0.
• Robert Jedrzejczyk (http://prog.olsztyn.pl/paslibvlc) for his PasLibVCLPlayer which enables WordSmith to play video.

Researchers from many other countries have also acted as alpha-testers and beta-testers and I thank them for their patience and feedback. I am also grateful to Nell Scott and other members of my family who have always given valuable support, feedback and suggestions.

Mike Scott

Feel free to email me at my contact address 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 which will run WordSmith unattended.

See also: custom processing
11.3 bibliography


Scott, Mike & Chris Tribble, 2006, Textual Patterns: keyword and corpus analysis in language education, Amsterdam: Benjamins.


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. http://lexically.net/wordsmith/version6/faqs/updating_or_reinstalling.htm explains.

Otherwise you should get a report popping up, giving "General" information about your PC and "Details" about the fault. This information will help me to fix the problem and will be saved in a small text file called wordsmith.elf, concord.elf, wordlist.elf, etc. When you quit the program, you will be offered a chance to email this to me.

The first thing you'll see when one of these happens is something like this:

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, retrieve your saved results 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

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, troubleshooting.

11.5 change language

If you have results computed with the wrong language setting, that can affect things, e.g. a key word listing depends on finding the words in the right order. To redefine the language of your data, choose Edit | Change Language, and in the resulting window

press Change once you have chosen a suitable alternative. If you choose a different one from the list of Alternatives, your Language and Text settings in the main Controller will change too. In this screenshot, pressing Change will change the language to Polish.

11.6 Character Sets

11.6.1 overview

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

If you are processing English only, your texts can be in ASCII, ANSI or Unicode; WordSmith handles both formats. If in other languages, read on...

To handle a text in a computer, programs need to know how the text is encoded. In its processing, the software sees only a long string of numbers, and these have to match up with what you and I can recognise as "characters". For many languages like English with a restricted alphabet, encoding can be managed with only 1 "byte" per character. On the other hand a language like Chinese, which draws upon a very large array of characters, cannot easily be fitted to a 1-byte system. Hence the creation of other "multi-byte" systems. Obviously if a text in English is encoded in a multi-byte way, it will make a bigger file than one encoded with 1 byte per character, and this is slightly wasteful of disk and memory space. So, at the time of writing, 1-byte character sets are still in very widespread use. UTF-8 is a name for a multi-byte method, widely used for Chinese, etc.
In practice, your texts are likely to be encoded in a Windows 1-byte system, older texts in a DOS 1-byte system, and newer ones, especially in Chinese, Japanese, Greek, in Unicode. What matters most to you is what each character looks like, but WordSmith cannot possibly sort words correctly, or even recognise where a word begins and ends, if the encoding is not correct. WordSmith has to know (or try to find out) which system your texts are encoded in. It can perform certain tests in the background. But as it doesn't actually understand the words it sees, it is much safer for you to convert to Unicode, especially if you process texts in German, Spanish, Russian, Greek, Polish, Japanese, Farsi, Arabic etc.

Three main kinds of character set, each with its own flavours, are Windows, DOS, and Unicode.

**Tip**

To check results after changing the code-page, select Choose Texts and View the file in question. If you can't get it to look right, you've probably not got a cleaned-up plain text file but one straight from a word-processor. In that case, take it back into the word-processor (see here 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, Choosing Accents & Symbols, Accented characters, Choosing Language

### 11.6.2 accents & symbols

When entering your search-word 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 by clicking.
Below, you will see which character has been selected

\[\text{ARABIC LETTER JEH FINAL FORM (FD6D)(Arial)}\]

with the current font (which affects which characters can be seen). You can choose a number of characters and then paste them into Concord, by right-clicking and choosing from the popup-menu:

- Paste α
- Paste ß
- Paste γ
- Paste δ
- Paste Ε
- Paste Ζ
- Paste Ω
- Paste Π
- Paste Σ
- Paste Τ
- Paste Υ

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, Change Language
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 (or printing 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.

![Concordance Example](image1)

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:

![Concordance in MS Word](image2)

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™. 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
avoid the heading and numbers in WordList or KeyWords too?
See advanced clipboard settings.

11.8 contact addresses

Downloads
You can get a more recent version at our website. There are also some free extra downloads (programs, word lists, etc.) there too. And links to sources of free text corpora.

Screenshots
Visit http://lexically.net/wordsmith/support/get_started_guides.html for screenshots of what WordSmith Tools can do. This may give you useful ideas for your own research and will give you a better idea of the limitations of WordSmith too!

Purchase
Visit http://lexically.net/wordsmith/purchasing.htm for details of suppliers.

Complaints & Suggestions
Best of all, join Google Groups WordSmith Tools group and post your idea there so others can see the discussion. Or email me (mike (at) lexically.net). Please give me as full a description of the problem you need to tackle as you can, and details of the equipment too. Please don't include any attachments over 200K in size. I do try to help but cannot promise to...

11.9 date format

Date Format
Japanese date format year_month_day_hour_minute. At least it is logical, going from larger to smaller. Why aren't URLs organised in a logical order too?

11.10 Definitions

11.10.1 definitions

valid characters
Valid characters include all the characters in the language you are working with which are defined (by Microsoft) as "letters", plus any user-defined acceptable characters to be included within a word (such as the apostrophe or hyphen). 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 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.
The term *token* is used to refer to running words and *type* to different words. So in *This is my book, it is interesting* we have 7 tokens but only 6 different types because *is* gets repeated.

**clusters**
A cluster is a group of words which follow each other in a text. The term phrase is not used here because it has technical senses in linguistics which would imply a grammatical relation between the words in it. In WordList cluster processing or Concord cluster processing there can be no certainty of this, though clusters often do match phrases or idioms. See also: general cluster information.

**sentences**
The sentence is defined as the full-stop, question-mark or exclamation-mark (.?!) and (equivalents in languages such as Arabic, Chinese, etc.) immediately followed by one or more word separators 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 or Viewer & Aligner technical information.)

**paragraphs**
Paragraphs are user-defined. See Starts and Ends of Text Segments for further details.

**headings**
Headings are also user-defined -- see Starts and Ends of Text Segments.

**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, Keyness, Key key-word, Associate.
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 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, Version Information, Contact Addresses.

11.12 drag and drop

You can get WordSmith to compute some results simply by dragging.

If you have WordList open you can simply drag a text file onto it from Windows Explorer and it will create a word-list there and then using default settings. Or if it is not open, drag your text file to the WordList6.exe file. Here, Hamlet is being dragged onto the WordList tool.

If you have KeyWords open you can simply drag a text file onto it from Windows Explorer. If you have a valid word list set as the reference corpus, it will compute the key words. Or if it is not open, drag your text file to the KeyWords6.exe file, as in this screenshot where the Dickens novel Dombey and Son.txt is being dragged onto the KeyWords file.

If you drag a word-list made by WordList (.LST ending), a concordance (.CNC), a key word list (.KWS) etc. onto the Controller it will open it with the appropriate tool.
11.13 edit v. type-in mode

Most windows allow you to press keys either
• to edit your data (edit mode), or
• to get quickly to a place in a list (type-in mode).
Concordance windows use key presses also for setting categories for the data, or for blanking out the search word.

In type-in mode, your key-presses are supposed to help you get quickly to the list item you’re interested in, e.g by typing theocr to get to (or near to) theocracy in a word list. If you’ve typed in 5 letters and a match is found, the search stops.

Changing mode is done in the menu: Settings | Typing Mode:

See also: user-defined categories.

11.14 file extensions

The standard file-extensions used in WordSmith are

<table>
<thead>
<tr>
<th>Extension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>.cnc</td>
<td>concordance file</td>
</tr>
<tr>
<td>.lst</td>
<td>word list</td>
</tr>
<tr>
<td>.mut</td>
<td>mutual information list</td>
</tr>
<tr>
<td>.dcl</td>
<td>detailed consistency list</td>
</tr>
</tbody>
</table>
.tokens, .types    word list index file
.kws               key words file
.kdb               key word database file
.base_pairs, .bas  WSCconcgram files
.e_index_cg        aligner list
.ali               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 than has been saved for each line, it'll need to re-read the source text. For KeyWords to calculate a dispersion plot, it needs to look at the source text to find out which KWs came near each other and compute positions of each KW in the text and KW links.

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

See also: Source texts, Editing filenames, Choosing source files, find files.

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, 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 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 in wordsmith6.ini.
• Concordance Folder: for your concordance files.
• KeyWords Folder: for your key-word list files.
• WordList Folder: where you will usually save your word-list files.
• Aligner: for your dual-text aligned work.
• Texts Folder: where your text files are to be found.
• Downloaded Media: where your sound & video files will be stored after downloading the first time from the Internet.
• Settings: where your settings files (.ini files and some others) are kept.

If you write the name of a folder which doesn't exist, WordSmith Tools will create it for you if possible. (On a network, this will depend on whether you have rights to create folders and save files.)

If you change your Settings folder, you should let WordSmith copy any .ini and other settings files which have been created so that it can keep track of your language preferences, etc.

Note: in a network, drive 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, 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.
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 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.

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 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**

\[(J - E) \text{ divided by the square root of } (E \times (1-P))\]

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**

\[(J \times 2) \text{ divided by } (F1 + F2)\]

where
J = joint frequency
F1 = frequency of word 1 or corpus 1 word count
F2 = frequency of word 2 or corpus 2 word count
Ranges between 0 and 1.

**Log Likelihood**

based on Oakes p. 170-2.

\[2 \times (a \text{ Ln } a + b \text{ Ln } b + c \text{ Ln } c + d \text{ Ln } d)
- (a+b) \text{ Ln } (a+b)
- (a+c) \text{ Ln } (a+c)
- (b+d) \text{ Ln } (b+d)
- (c+d) \text{ Ln } (c+d)
+ (a+b+c+d) \text{ Ln } (a+b+c+d)\]

where
a = joint frequency
b = frequency of word 1 - a
c = frequency of word 2 - a
d := frequency of pairs involving neither word 1 nor word 2
and "Ln" means Natural Logarithm

See also: this link from Lancaster University, Mutual Information
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

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.
11.22 international versions

WordSmith can operate with a series of interfaces depending on the language chosen.

If you choose French this is what you see in all of WordSmith.

See also: acknowledgements
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 and b) patience.

You can have as many copies of each Tool running at any one time as you like. Each one allows you to work on one set of data. Tags to ignore or ones containing an asterisk can span up to 1,000 characters.

When searching for tags to determine whether your text files meet certain requirements, only the first 2 megabytes of text are examined. For Ascii that's 2 million characters, for Unicode 1 million.

Tip
Press F9 to see the “About” box -- it shows the version date and how much memory you have available. If you have too little memory left, try a) closing down some applications, b) closing WordSmith Tools and re-entering.

See also: Specific Limitations of each Tool

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, though you can specify an unlimited number of concordance search-words in a search-word file.
Each concordance can store an unlimited number of collocates with a maximum horizon of 25 words to left and right of your search-word.

WordList limitations
A head entry can hold thousands of lemmas, but you can only join up to 20 items in one go using F4. Repeat as needed. Detailed Consistency lists can handle up to 50 files.

KeyWords limitations
One key-word plot per key-word display. (If you want more, call up the same file in a new display window.)
number of link windows per key-word plot display: 20.
number of windows of associates per key key-word display: 20.

File Utilities: Splitter limitations
Each line of a large text file can be up to 10,000 characters in length. That is, there must be an <Enter> from time to time!
**Text Converter limitations**

There can be up to 500 strings to search-and-replace for each.
Each search-string and each replace-string can be up to 80 characters long.
An asterisk must not be the first or last character of the search-string.
When the asterisk is used to retain information, the limit is 1,000 characters.

**Viewer & Aligner limitations**

If you choose the View option when choosing texts, Viewer & Aligner will call up the first 10 source text files selected.

When choosing texts or jumping into the middle of a text (e.g. after choosing in Concord), Viewer & Aligner will only process 10,000 characters of each file, to speed things up in the case of very large files, but you can get it to "re-read" the file by pressing to refresh the display, after which it will read the whole text.

See also: General Limitations

**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 (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 http://lexically.net) you can click to access your browser.

**Links between the various Tools**

The programs in WordSmith Tools are linked to each other via wordsmith.exe (the one which says " WordSmith Tools Controller" in its caption, and is found in the top-left corner of your screen). This handles all the defaults, such as colours, folders, fonts, stop lists, etc.

In general, if you press 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, or view a dispersion plot.

**Example:** after computing a key words database using 300 business texts, you discover that
the word bid seems to be a key key-word, and that it’s associated with company, shares etc. Place the highlight on bid, press Control-C and a concordance will be computed using the same 300 texts. Now you can check out the contexts: is bid a bid for power, or is it part of a tendering process?

Example: you have a concordance of green. Now press Control-W to generate a word list of the same text files. Press Control-K to compare this word list with a reference corpus list to see what the key words are in these text files.

11.25 keyboard shortcuts

scrolling windows:

Control+Home  to top of scrollable list
Control+End  to end of scrollable list
if it’s ordered  type-in your search-word
alphabetically:  
and if it scrolls  Home -- to left edge
horizontally:  End -- to right edge

hotkeys:

Shift-cursor keys  block a section

F1  help ?
Ctrl+F2  save results save
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 re-sort
Ctrl+F6  reverse word sort
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
**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 machine, especially if there's plenty of RAM.

You can run WordSmith from a memory stick on a fast computer better than on a slow computer. (You can run WordSmith on a tiny 10" screen laptop with Windows Starter and little power but all applications on those are slow and there is not much screen for your results.)

There is no Apple Mac version but see [http://lexically.net/wordsmith/mac_intel.htm](http://lexically.net/wordsmith/mac_intel.htm) for details on how to use WordSmith on a Mac.

**11.27 manual for WordSmith Tools**

This help file exists in the form of a manual, which you get when you install. 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.
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.

- **auto-join**
  joins (lemmatises) automatically.

- **auto-size**
  re-sizes each line of a display so that each one shows as much data as it should. Most windows have lines of a fixed size but some, e.g. in Viewer, allow you to adjust row heights. This adjusts line heights according to the current highlighted column of data.

- **close (Ctrl+W)**
  closes a window of data

- **clumps**
  computes clumps in a keywords database

- **regroup clumps**
  regroups the clumps

- **clusters**
  computes concordance.

- **collocates**
  shows collocates using concordance data.

- **compute**
  calculates a new column of data based on calculator functions and/or existing data.

- **redo collocates**
  recalculates collocates, e.g. after you've deleted concordance lines.

- **column totals**
  computes totals, min, max, mean, standard deviation for each column of numerical data.

- **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 your data to a variety of different places (the printer, a text file, the clipboard, etc.).

- **edit**
  allows editing of a list or searches for a word.

- **exit (Alt+X)**
quits a Tool.

1 edit or type-in mode
alternates between edit and type-in mode.

fill
opens a new window showing the file names from which the current data derived. If necessary you can edit them.

F 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).

layout
This allows you to alter many settings for the layout: the colour of each column, whether to hide a column of data, typefaces and column widths.

links
computes between words in a key-words plot.

mark
marks an entry for joining or finding files.

== match lemmas
checks each item in the list against ones from a text file of lemmatised forms and joins any that match.

== match list
matches up the entries in the current list against ones in a "match list file" or template, marking any found with (~).

relation
computes mutual information or similar scores in a WordList index list.

new... (Ctrl+N)
gets you started 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.

^ play media (Ctrl+M)
plays a media file.

plot
opens a new window showing a Concord dispersion plot or KeyWords plot.
print preview (F3)
previews your window data for printing (Ctrl+P); can print to file, which is equivalent to "save as text"

redo
undoes an undo.

refresh (F5)
re-draws the screen (in Viewer re-reads your text file).

remove duplicates
removes any duplicate concordance lines.

replace
search & replace, e.g. to replace drive or folder data, when editing file-names where the source
texts have been moved.

re-sort
re-sorts lists (e.g. in frequency as opposed to alphabetical order) in Concord, KeyWords or WordList.

ruler
shows/hides vertical divisions in any list; text divisions in a KeyWords plot. Click ruler in a menu to
turn on or off or change the number of ruler divisions for a plot.

save (also Ctrl+F2)
saves your data using existing file-name; if it's a new file asks for file-name first.

save as
saves after asking you for a file-name.

save as text
saves as a .txt file: plain text.

search
searches within a list.

shrink
reduces the height of all rows to a smaller fixed height. See grow above.

statistics
shows detailed statistics.

statusbar
toggles on & off the "status bar" (at the bottom of a window, shows comments and the status of what has been done).

summary statistics
opens a new window showing summary statistics, e.g. proportion of lemmas to word-types.

toolbar
toggles on & off a toolbar with the same buttons on it as the ones you chose when you customised
popup menus.

undo (Ctrl+U)
undoes last operation.
unjoin
unjoins any entries that have been joined, e.g. lemmatised entries.

view source text
shows the source text and highlights any words currently selected in the list.

Microsoft Excel or Word™
save formatted data for Excel or Word.

wordlist
within KeyWords, makes a word list using the current data.

zap (Ctrl+Z)
zaps any deleted entries.

see also: Keyboard Shortcuts, Customising popup menus.

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.

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):

Warning: Saving as a text file will cause all formatting, pictures, and objects in your file to be lost.

Text encoding:
- [ ] Windows (Default)
- [ ] MS-DOS
- [ ] Other encoding: Unicode (2-bytes)

Options:
- [ ] Insert line breaks
- [ ] End lines with: CR / LF
- [ ] Allow character substitution

Preview:

10.99, Today Programme, BBC. ¶

James Naughtie asked a Conservative spokesman whether Michael Portillo’s chances of future leadership of the Conservative Party might have been affected by recent news concerning his sexuality. The spokesman stated his reply saying there was no question of a leadership contest, William Hague was fully supported by all including Michael Portillo. ¶

---

Naughtie: That wasn’t the question. ¶

Spokesman: I realise that. But I’m giving you the answer. ¶
11.30 never used WordSmith before

For users who are starting out with WordSmith for the first time, the whole process can seem complex. (After all, the first time you used word-processing software that seemed tricky -- but you already knew what a text is and how to write one...)

So a small text file accompanies the WordSmith installation, and if WordSmith thinks you have never used it before, it will automatically choose that text file for you to start using Concord, WordList etc. WordSmith's method of knowing that you are a new user is

1) have any concordances or wordlists been saved?

and

2) has no set of favourite text files been saved for easy retrieval?

11.31 numbers

Depending on Language and Text Settings, you might wish to include or exclude numbers from word lists.

11.32 plot dispersion value

The point of it

A dispersion value is the degree to which a set of values are uniformly spread. Think of rainfall in the UK -- generally fairly uniformly spread throughout the year. Compare with countries which have a rainy season.

In linguistic terms, one might wish to know how the occurrences of a word like skull are distributed in Hamlet, and WordSmith has shown this in plot form since version 1. The dispersion value statistic gives mathematical support to this and makes comparisons easier.

How it is calculated

The plot dispersion calculated in KeyWords and Concord dispersion plots uses the first of the 3 formulae supplied in Oakes (1998: 190-191), which he reports as having been evaluated as the most reliable.

Like the ruler, it divides the plot into 8 segments for this.

It ranges from 0 to 1, with 0.9 or 1 suggesting very uniform dispersion and 0 or 0.1 suggesting "burstiness" (Katz, 1996)

See also: KeyWords plot, Concord dispersion plot.
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, 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

11.34 reference corpus

Reference Corpus

A corpus of text which you use for comparative purposes. For example, you might want to compare a given piece of text with the British National Corpus, a collection of 100 million words. Useful when computing key words.

In the Controller you can set your reference corpus word list for KeyWords and Concord to make use of. (That is, a word list created using the WordList tool.)

11.35 restore last file

By default, the last word list, concordance or key words listing that you saved or retrieved will be automatically restored on entry to WordSmith Tools. If the last Tool used is Concord, a list of your 10 most recent search-words will be saved too. This feature can be turned off temporarily via a menu option or permanently in wordsmith6.ini (in your Documents\wsmith6 folder).
11.36 selecting multiple entries

To select more than one entry in a word-list, concordance, key word list etc, hold down Control first, and next select all the rows you are interested in. To mark entries for joining in lemmatisation, you can choose Edit | Mark (F5) in the menu.

For example, to do a search from a word-list of these items, I held down Control and pressed FEB, FEBRUARY, FEBUARY and FEBURARY, then chose Edit | Concordance.

The resulting concordance shows the last two entries are indeed mis-spellings.

To clear the marking, click the list and press Control.
11.37 single words v. clusters

The point of it...
Clusters are words which are found repeatedly together in each others' company, in sequence. They represent a tighter relationship than collocates, more like multi-word units or groups or phrases. (I call them clusters because groups and phrases already have uses in grammar and because simply being found together in software doesn't guarantee they are true multi-word units.) Biber calls them "lexical bundles". Language is phrasal and textual. It is not helpful to see it as a matter of selecting a word to fill a grammatical "slot" as implied by structural theories. Words keep company: the extreme example is idiom where they're bound tightly to each other, but all words have a tendency to cluster together with some others. These clustering relations may involve colligation (e.g. the relationship between depend and on, collocation, and semantic prosody (the tendency for cause to come with negative effects such as accident, trouble, etc.). WordSmith Tools gives you two opportunities for identifying word clusters, in WordList and Concord. They use different methods. Concord only processes concordance lines, while WordList processes whole texts.

How 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
time
(note not "time there" because of the comma)
there was (etc.)
With a three-word cluster setting, it would send

Once upon a
time
there was a
beautiful princess
But 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 ;,.!?

See also: concgrams.

11.38 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. 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

<table>
<thead>
<tr>
<th>search-word</th>
<th>source</th>
<th>speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>quickly</td>
<td>CD-ROM</td>
<td>6 million words per minute</td>
</tr>
<tr>
<td>quickly</td>
<td>hard disk</td>
<td>12 million wpm</td>
</tr>
<tr>
<td>the</td>
<td>CD-ROM</td>
<td>900,000 wpm</td>
</tr>
<tr>
<td>the</td>
<td>hard disk</td>
<td>1 million wpm</td>
</tr>
<tr>
<td>thez</td>
<td>CD-ROM</td>
<td>6 million wpm</td>
</tr>
<tr>
<td>thez</td>
<td>hard disk</td>
<td>16 million wpm</td>
</tr>
</tbody>
</table>

Tests using a set of text files ranging from 20K down to 4K, using quickly as the search-word, gave speeds of 2 million wpm rising with the longer files to 4 million wpm. Making a word list on the same set of files gave an average speed of 800,000 wpm. On the 55MB text file the speed was around 1.35 million wpm.

These data suggest that factors which slow concordancing down are, in order, word rarity (the was much slower than quickly or the non-existent thez), text file size (very small files of only 500 words or so (3K) will be processed about three times as slowly as big ones) and disk speed (the outdated quad speed CD-ROM being roughly half the speed of the 12ms hard disk). When Concord finds a word it has to store the concordance line and collocates and show it (so that you can decide to suspend any further processing if you don't like the results or have enough already). This is a major factor slowing down the processing. Second, reading a file calls on the computer's file management system, which is quite slow in loading it, in comparison with Concord actually searching through it. Third, disk speeds are quite varied, floppy disks being much the worst for speed.

If processing seems excessively slow, close down as many programs as possible and run WordSmith Tools again. Or install more RAM. Get advice about setting Windows to run efficiently (virtual memory, disk caches, etc.) Use a large fast hard drive.

You can run other software while the programs are computing, but they will take up a lot of the processor's time. Shoot-em-up games may run too jerkily, but printing a document at the same time should be fine.

11.39 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.
11.40 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.

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.
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.

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.

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, Dispersion Plots, Acknowledgements

11.41 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 available. If you have a demonstration version this will be stated immediately below your name.

Check the date in this box, which will tell you how up-to-date your current version is. As suggestions are incorporated, improved versions are made available for downloading. Keep a copy of your registration code for updated versions.

You can click on the WordSmith graphic in the About box to see your current code.
11.41.1 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 handling including Tag Concordancing
- converter for previous data
- zip file handling
- easier exporting of data to Microsoft Word and Excel
- Unicode text handling, allowing more languages to be processed
- possibility of altering the data as it comes in, e.g. for language-specific lemmatisation
- the old limitations of 16,000 lines of data went. (The theoretical limit for a list of data is over 134 million lines.)

See also: What's New in the current version, Contact Addresses.
11.42 zip files

Zip files are files which have been compressed in a standard way. WordSmith can now read and write to .zip files.

The point of it...

Apart from the obvious advantage of your files being considerably smaller than the originals were, the other advantage is that less disk space gets wasted like this: any text file, even a short one containing on the word "hello", will take up on your disk something like 4,000 bytes or maybe up to 32,000 depending on your system. If you have 100 short files, you would be losing many thousands of bytes of space. If you "zip" 100 short files they may fit into just 1 such space. Zip files are used a lot in Internet transmissions because of these advantages. If you have a lot of word lists to store, it will be much more efficient to store them in one .zip file.

The "cost" of zipping is a) the very small amount of time this takes, b) the resulting .zip file can only be read by software which understands the standard format. There are numerous zip programs on the market, including PKZip™ and Winzip™. If you zip up a word list, these programs can unzip it but won't be able to do anything with the finished list. WordSmith can first unzip it and then show it to you.

How to do it...

Where you see an option to create a zip file, this can be checked, and the results will be stored where you choose but in zipped form with the .zip ending. If you choose to open a zipped word list, concordance, text file, etc. and it contains more than one file within it, you will get a chance to decide which file(s) within it to open up. Otherwise the process will happen in the background and will not affect your normal WordSmith processing.
12 Troubleshooting

12.1 list of FAQs

See also: logging

These are the Frequently Asked Questions.

There's a much longer list of explanations under Error Messages.

Can't process apostrophes

Is this Russian, Greek or English? strange symbols in display

It crashed

It doesn't even start

It takes ages

Keys don't respond

Line beyond demo limit

Mismatch between Concord and WordList results

No tags visible in concordance

Printing problem

Text is unreadable because of the colours

Too much or too little space between columns

Wordlist out of order

Won't slice pineapples

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, or else replace them in your text files using Text Converter.

See also: settings

12.3 column spacing

column spacing is wrong

You can alter this by clicking on the layout 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 has the same format. For example, if Text to Ignore has "", any tags such as <title>, <quote>, etc. will be cut out of the concordance unless you specify them in a tag file.

Solution: specify the tag file and run the concordance again.
12.5 Concord/WordList mismatch

Concord/WordList mismatch

If WordList finds a certain number of occurrences of a (word list) cluster but Concord finds a different number, this is because the procedures are different. WordList proceeds word by word, ignoring punctuation (except for hyphens and apostrophes). When Concord searches for a (concordance) cluster it will (by default) take punctuation into account: you can change that in the settings 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. 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 and a full version is: with the latter you can see or print all the data, with the former you'll be able to see only about 25 lines of output.

12.8 funny symbols

weird symbols

funny symbols when using WordSmith Tools

1. Check your text files. Look at them in Notepad. Do they contain lots of strange symbols? These may be hidden codes used by your usual word-processor. Solution: open them in your usual word-processor and Save As, with a new name, in plain text format, sometimes called "Text Only" or .txt. In Word 2003 the option looks like this:

   ![Save as Type: Plain Text (*.txt)](image)

   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 to clean up and convert 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 or Text Characteristics, the lists will all be updated to show the current font and character set, unless you first minimize any window which would be affected.

**funny symbols when reading WordSmith data in another application**

WordSmith Tools can Save or Save As and Saves as text by printing to a file. “Save” and “Save As” will store the file in a format for re-use by 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.

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 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

Words are sorted according to Microsoft routines which depend on the language. If you process Spanish but leave the Language settings to "English", you will get results which are not in correct Spanish order, (e.g. LL will come just before LM).
Solution: choose your language and re-compute the word-list.
Error Messages

Section XIII
13 Error Messages

13.1 List of Error Messages

List of Error Messages
See also: Troubleshooting.

- Can only save WORDS as ASCII
- Can't call other Tool
- Can't make folder as that's an existing filename
- Can't merge list
- Can't read file
- Character set reset to <x> to suit <language>
- Concordance file is faulty
- Concordance stop list file not found
- Conversion file not found
- Destination folder not found
- Disk problem: File not saved
- Dispersions go with concordances
- Drive not valid
- Failed to access Internet
- Failed to create new folder name
- File access denied
- File contains none of the tags specified
- File not found
- Filenames must differ
- Full drive: folder name needed
- function not working properly yet
- INI file not found
- Invalid Concordance file
- Invalid file name
- Invalid Keywords Database file
- Invalid Keywords file
- Invalid Wordlist Comparison file
- Invalid Wordlist file
- Joining limit reached: join & try again
- Key words file is faulty
- Keywords Database file is faulty
- Limit of 500 file-based search-words reached
- Links between Tools disrupted
- Match list details not specified
- Must be a number
- Network registration running elsewhere or vice-versa
- No access to text file: in use elsewhere?
- No associates found
- No clumps identified
- No clusters found
- No collocates found
- No concordance entries found
- No concordance stop list words
- No deleted lines to Zap
- No entries in Keywords Database
- No Key Words found
455  WordSmith Tools Manual

No key words to plot
No keyword stop list words
No lemma list words
No match list words
No room for computed variable
No statistics available
No stop list words
No such file(s) found
No tag list words
Not a valid number
No wordlists selected
Only X% of reference corpus words found
Original text file needed but not found
Registration string is not correct
Registration string must be 20 letters long
Short of Memory!
Source Folder file(s) not found
Stop list file not found
Stop list file not read
Tag file not found
Tag list file not read
This function is not yet ready!
This is a demo version
This program needs Windows 95 or greater
To stop getting this annoying message, Update from Demo in setup.exe
Too many ignores (50 limit)
Too many sentences (8000 limit)
Two files needed
Truncating at xx words -- tag list file has more!
Unable to merge Keywords Databases
Why did my search fail?
Word list file not found
Wordlist comparison file is faulty
Word-list file is faulty
WordSmith Tools has expired: get another
WordSmith Tools already running
WordSmith version mis-match
xx days left

13.2  .ini file not found

.ini file not found

On starting up, WordSmith looks for the `wordsmith6.ini` file which holds your current defaults. If you've removed or renamed it, restore it. This file should be in a sub-folder of your Documents folder called `wsmith6`.

13.3  administrator rights

administrator rights

If you see this error message it's because you need Administrator rights to register WordSmith. Try searching for "Run as Administrator" or this link.
13.4 base list error

base list error
WordSmith is trying to access an word or concordance line above or below the top or bottom of the data computed. This is a bug.

13.5 can only save words as ASCII

Can only save WORDS as Plain Text
WordSmith Tools can't save graphics as a text file. If you get this error message, you can only save this type of data by copying to the clipboard 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 your work, first. Then, if necessary, close down WordSmith Tools altogether, then start the main wordsmith6.exe program again.

13.7 can't make folder as that's an existing filename

Can't make folder as that's an existing filename
If you already have a file called C:\TEMP\FRED, you can't make a sub-folder of C:\TEMP called FRED. Choose a new name.

13.8 can't compute key words as languages differ

Can't compute key words as languages differ
Key words can only be computed if both the text file and the reference corpus are in the same primary language. You can compute KWs using 2 different varieties of English or 2 different varieties of Spanish, but not between English and French.

13.9 can't merge list with itself!

Can't merge list with itself
You can only merge 1 word list or key word database with 1 other at a time. Select (by clicking while holding down the Control key) 2 file-names in the list of files.

13.10 can't read file

Can't read file
If this happens when starting up WordSmith Tools, there is probably a component file missing. One example is sayings.txt, which holds sayings that appear in the main Controller 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 and languages than it does now. Accordingly, data saved in the format used before that version may not "know" what language it was based on. If you get this message when opening up an old WordSmith data file, it's because WordSmith doesn't know what language it derived from. Through gross linguistic imperialism, it will by default assume that the language is English!

If the data are okay, just click the save button so that next time it will "know" which language it's based on. If not, reset the language to the one you want in the Controller, Language Settings | Text, then re-save the list.

13.12 concordance file is faulty

Concordance file is faulty

Each type of file created by WordSmith Tools has its own default filename extension (e.g. .CNC, .LST) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! WordSmith has detected that the file you're calling up wasn't produced by the current version of Concord.

13.13 concordance stop list file not found

Concordance stop list file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.14 confirmation messages: okay to re-read

Okay to re-read?

A confirmation message. To proceed, Viewer & Aligner will now re-read the disk file. This will affect any alterations you've already made to the display. You may wish to save first and then try again later.

Also, Viewer & Aligner will try to read the whole text file. If you have a very big file on a slow CD-ROM drive, this will take some time.

13.15 conversion file not found

Conversion file not found

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.
13.16 destination folder not found
   Destination folder not found
   WordSmith couldn't find that folder; perhaps it's mis-spelt.

13.17 disk problem -- file not saved
   Disk problem: File not saved
   Something has gone wrong with a disk writing operation. Perhaps there's not enough room on the drive. If so, delete some files on that drive.

13.18 dispersions go with concordances
   Dispersions go with concordances
   They can't be saved separately.

13.19 drive not valid
   Drive not valid
   WordSmith is unable to access this drive. This could happen if you attempt to access a disk drive which doesn't exist, e.g. drive P: where your drives include A:, C:, D: and E:.

13.20 failed to access Internet
   Failed to access Internet
   This function relies on a) your having an Internet browser on your computer, b) your system "associating" an Internet URL ending .htm with that browser.

13.21 failed to create new folder name
   Failed to create new folder or file-name
   A folder and a file cannot have the same name. If you already have a file called C: \ TEMP \ FRED, you can't make a sub-folder of C: \ TEMP called FRED. Choose a new name. Or you don't have rights to create files in that folder. Or something went wrong while WordSmith was trying to write a file, for example the disk was full up.

13.22 failed to read file
   Failed to Read
   This may have happened
   a) because you included a text file which happens to be empty (zero size), or
   b) because your disk filing system has got screwed up, which is especially likely to occur if you often use large files in your word processor (in which do a disk cleanup) or
   c) because you tried to use the wrong kind of file for the job (for example the KeyWords procedure won't work if you choose text files as your word-lists).
13.23 failed to save file

Failed to Save

Maybe because you had the same file open in another program or another instance of the Tool you're running. If so, close it and try again.
Or because the folder you're saving to is a read-only folder on a network, or because the disk is full, or because your disk filing system has got screwed up. This last problem is quite common, actually, and is especially likely to occur if you often use large files in your word processor. In that case run Programs | Accessories | System Tools | Disk Defragmenter.
If you're working on a network, you will be able to save on certain drives and folders but not others; the solution is to try again on a memory stick or a hard disk drive which you do have the right to save to.

13.24 file access denied

File Access Denied

Maybe the file you want is already in use by another program. You'll find most word-processors label any text files open in them as "in use", and won't let other programs access them even just to read them. Close the text file down in your word processor.

13.25 file contains none of the tags specified

You specified tags, but none of them were found.

13.26 file has "holes"

File has "holes"

Text files are supposed to contain only characters, punctuation, numbers, etc. without any unrecognised ones such as character(0). The problem could have arisen because it was transferred from one system to another, part of the disk is corrupted, or else maybe the file contains unrecognised graphics (or else it is not a plain text file but e.g. a Word document).

You can solve this problem by converting the text using the Text Converter. If it is a plain text with holes these will be replaced by spaces. You can find texts with holes using the File Utilities.

13.27 file not found

File not found

This message, like Original Text not found, can appear when WordSmith needs to access the original source text used when a list was created, but cannot find it. Have you deleted or moved it? If the file is still available, you may be able to edit the file names in the file name window of this list.
Or the message may come after you've supplied the file name yourself. You may have mis-typed it. If typing in a file name, remember to include the full drive and folder as well as the file name itself.
13.28 filenames must differ!

Filenames must differ
You can't compare a file with itself.

13.29 folder is read-only

folder is read-only
For some purposes, WordSmith needs to save files e.g. lists of results you have made so that you can get at recent files again. To do this it needs a place where your network or operating system lets you save. Usually \wsmith6 is fine, but in some institutional settings the drive or folder may be "read-only". If you see this message, choose Folder Settings and select there a folder where you can write as well as read.

13.30 for use on X machine only

For use on pc named XXX only
The software was registered for use on another PC. If you get this message, please re-install as appropriate.

13.31 form incomplete

Form incomplete
You tried to close a form where one or more of the blanks needed to be filled in before WordSmith could proceed.

13.32 full drive & folder name needed

Full drive::folder name needed
When typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.33 function not working properly yet

function not working properly yet
This is a function under development, still not fully implemented.

13.34 invalid concordance file

Invalid Concordance file
Each type of file created by WordSmith Tools has its own default filename extension (e.g. .CNC, .LST) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! WordSmith has detected that the file you're calling up wasn't produced by the current version of Concord.
13.35 invalid file name

Invalid file name
File names may not contain spaces or certain symbols such as ? and *.

13.36 invalid KeyWords database file

Invalid Keywords Database file
Each type of file created by WordSmith Tools has its own default filename extension (e.g. .KWS, .KDB) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .KDB file to .TXT, or vice-versa! WordSmith has detected that the file you're calling up wasn't produced for a database by the current version of KeyWords.

13.37 invalid KeyWords calculation

Invalid Keywords calculation
For KeyWords to calculate the key-words in a text file by comparing it with a reference corpus, both must be in the same language, both must be sorted in the same way (alphabetical order, ascending) and they should both be in the same format (Unicode or single-byte). If you see this message you are trying to compute KWs without meeting these criteria. Solution: open each word-list and check to see if it is OK and that it is sorted alphabetically in the same way (in the Alphabetical view, click the top bar to re-sort in ascending alphabetical order), then save it. Check they have both been made with the same language & format settings and if necessary re-compute one or both of them.

13.38 invalid WordList comparison file

Invalid Wordlist Comparison file
Each type of file created by WordSmith Tools has its own default filename extension (e.g. .LST, .CNC) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .CNC file to .TXT, or vice-versa! WordSmith has detected that the file you're calling up wasn't produced as a comparison file by WordList.

13.39 invalid WordList file

Invalid Wordlist file
Each type of file created by WordSmith Tools has its own default filename extension (e.g. .LST, .CNC) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a .LST file to .TXT, or vice-versa! WordSmith has detected that the file you're calling up wasn't produced by the current version of WordList.

13.40 joining limit reached

Joining limit reached: join & try again
Only a certain number of words can be lemmatised in one operation. If you reach the limit and get this message,
1. lemmatise by pressing F4,
2. place the highlight on the head entry again
3. press F5 and carry on lemmatising by pressing F5 on each entry you wish to attach to the head entry
4. when you've done, press F4 to join them up.

13.41 KeyWords database file is faulty

**KeyWords database file is faulty**

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g., `.KDB`, `.KWS`) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a `.KDB` file to `.TXT`, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced for a database of keywords, by the current version of KeyWords.

13.42 KeyWords file is faulty

**KeyWords file is faulty**

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g., `.KWS`, `.KDB`) and its own internal structure. If you have another file with the same extension produced by another program, this will not be compatible. It would not be sensible to rename a `.KWS` file to `.TXT`, or vice-versa! **WordSmith** has detected that the file you're calling up wasn't produced by the current version of KeyWords.

13.43 limit of file-based search-words reached

**Limit of search-words reached**

No more than 15 search-words can be processed at once, unless you use a file of search words to tell Concord to do them in a batch, where the limit is 500.

13.44 links between Tools disrupted

**Links between Tools disrupted**

**WordSmith Tools Controller** or an individual Tool has tried to call another Tool and failed. There may have been a fault in another program you’re running or a shortage of memory. As inter-tool communication links are vital in this suite, you should exit WordSmith and re-enter.

13.45 match list details not specified

**Match list details not specified**

You pressed the **Match List** button but then failed to choose a valid match list file or else to type in a template for filtering. Try again.

13.46 must be a number

**Must be a number**

You typed in something other than a number. Be especially careful with lower-case L and 1, and O (the letter) instead of 0 (the number).
13.47 mutual information incompatible

Mutual information list is incompatible

A mutual information list derives from an index file, and knows which index file it derives from when computed. Normally when it opens up, it opens up the corresponding index file too. If that index file is not found on your PC or has been renamed, you will see this message. The mutual information can still be accessed but a) what you see in terms of Frequency and Alphabetical lists refers to a different index file, and b) it will not be possible to get concordances directly from the listing.

13.48 network registration used elsewhere

Network registration running elsewhere or vice-versa

The site licence registration for use on a network is not valid for use on a stand-alone pc, and vice-versa. If you get this message, please re-install as appropriate.

13.49 no access to text file - in use elsewhere?

No access to text file: in use elsewhere?

The file cannot be accessed. Perhaps another application is using it. If so, close down the file in that other application and try again.

13.50 no associates found

No associates found

Alter settings (Settings | Min & Max Frequencies) and try again.

13.51 no clumps identified

No clumps identified

Alter settings and try again.

13.52 no clusters found

No clusters found

Alter the settings (Settings | Clusters) and try again. There were too few concordance lines to find the minimum number needed, or the cluster length was too great.

13.53 no collocates found

No collocates found

In the Controller, alter the settings (Concord settings | Min. Frequency) and try again. There were too few concordance lines to find the minimum number needed.
13.54 no concordance entries

No concordance entries found

If you got no concordance entries, either a) there really aren't any in your text(s), b) there's a problem with the specification of what you're seeking, or c) there's a problem with the text selection. Check how you've spelt the search-word and context word. If you're using accented text [151], check the format of your texts. If you're using a search-word file [151], ensure this was prepared using a plain Windows word-processor such as Notepad.

Have you specified any wildcards [151] (" 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 [151] or two. You're more likely to find book* than book.

13.55 no concordance stop list words

No concordance stop list words

13.56 no deleted lines to zap

No deleted lines to Zap

You pressed Ctrl+Z but hadn't any deleted lines to zap. No harm done.

13.57 no entries in KeyWords database

No entries in Keywords Database

Alter settings and try again.

13.58 no fonts available

No fonts available for language

The operating system does not have a font which can show the characters for that language. You need to find and install a font.

13.59 no key words found

No Key Words found

Alter settings and try again. The minimum frequency is set too high and/or the p value [151] too small for any key words to be detected. For very short texts a minimum frequency of 2 may be needed.
13.60 no key words to plot

No key words to plot
Had you deleted them all?

13.61 no KeyWords stop list words

No keyword stop list words
WordSmith either failed to read your stop-list file or it was empty.

13.62 no lemma list words

No lemma match list words
WordSmith either failed to read your lemma list file or it was empty.

13.63 no match list words

No match list words
WordSmith either failed to read your match list file, or it was empty, or you forgot to check the action to be taken (one option is None). Or you tried to match up using a list of words, or a template, when the current column has only numbers. Or else there really aren't any like those you specified!

13.64 no room for computed variable

No room for computed variable
There isn't enough space for the variable you're trying to compute.

13.65 no statistics available

No statistics available
Some types of word list created by WordSmith Tools, e.g. a word list of a key words database have words in alphabetical and frequency order but no statistics on the original text files. You cannot therefore call the statistics up in WordList. You might also see this message if the statistics file you're trying to call up is corrupted.

13.66 no stop list words

No stop list words
WordSmith either failed to read your stop-list file or it was empty.

13.67 no such file(s) found

No such file(s) found
You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.
13.68 no tag list words

No tag list words
WordSmith either failed to read your tag file or it was empty.

13.69 no word lists selected

No word lists selected
For WordSmith to know which word lists to compare, you need to select them, by clicking on one in each folder. If you've changed your mind, press Cancel.

13.70 not a valid number

Not a valid number
Either you've just typed in, or else WordSmith Tools has just attempted to read (e.g. from wordsmith6.ini, the defaults file), something which is expected to be a number but wasn't. Computers will not see the capital O as equivalent to the number 0. Or else there is a number but accompanied by some other letters or symbols, e.g. £30. If this happens when WordSmith is starting up, check out the wordsmith6.ini file for mistakes.

13.71 not a WordSmith file

The file you are trying to open is not a WordSmith Tools file. WordSmith makes files containing your results, files whose names end in .LST, .CNC, .KWS, etc. These are in WordSmith's own format and cannot be opened up by Microsoft Word -- likewise a plain text file or a Word .doc cannot usually be read in by WordSmith as a data file, but only as a text file for processing.

See also: Converting Data from Previous Versions

13.72 not a current WordSmith file

Not a Current WordSmith File
The file you are trying to open was made using WordSmith but either
- it's a file made using version 1-3
  or
- it's a file made with the beta version of WordSmith and the format has had to change (sorry!)

If the former, you may be able to convert it using the Converter.

13.73 nothing activated

Nothing activated
Some forms have choices labelled "Activated" which you can switch on and off. If they are unchecked, you can still see what they would be but WordSmith will ignore them.
13.74 **Only X% of words found in reference corpus**

When WordSmith computes key words it checks to see that most of the words in your small word-list are found in the reference corpus, as would be expected. If less than 50% are found, you will get this warning. That is a bit unusual, and is supplied as a warning that for example there might be something strange about one of your two texts. If you know there is nothing strange, then you could ignore the message.

If you are processing clusters you are much more likely to see this warning, however, as the chance of 3-word strings matching in the two lists is less than that of single words matching.

It is up to you to decide whether there is some error in what you are doing or it is OK for many of your smaller word list's words/clusters not to be found in the reference corpus word list.

It might not be so unusual if your reference corpus was very small. But if it is indeed very small, the whole procedure is not very reliable. WordSmith simply looks at the frequencies of each word form and uses basic statistics to compute how greatly they differ in frequency. Basic statistics rely on a notion of what can be expected. If the reference corpus is incredibly small, WordSmith's computation of what is to be expected isn't really very reliable. As a dumb example if you met three citizens of a country you have never visited, and all looked fat, you might suppose the people of that country to be fat in general, but the sample size is not reliable for such an expectation. The KW procedure isn't really proof of anything, incidentally. Words don't occur in texts at all randomly and all ordinary basic statistics can do in my opinion is give us food for thought. So a KW listing isn't proof of anything but it may well give good ideas as to what may prove interesting avenues for research.

13.75 **original text file needed but not found**

To proceed, **WordSmith** needed to find the original text file which the list was based on. But it has been moved or renamed.

Or if on a network, your network connection is not mapped, or the network is down ...or else the right disk or CD-ROM is not in the drive!

13.76 **printer needed**

WordSmith needs a printer driver to be installed, even if you never actually print anything. You don't need to buy a printer or to switch a printer on, but the *Print Preview* function in Concord, WordList, KeyWords etc. does need to know what sort of paper size you would print to. If you get a message complaining that no printer has been installed, choose Start | Settings | Printers & Faxes and install a default printer (any printer will do) in Windows.

13.77 **registration code in wrong format**

**Registration code unexpectedly short**

PASTE the registration supplied into the box; only paste into the Name or Other Details boxes the details supplied.

If you see this message on registering you may have a registration for a previous major version. If so, contact sales at lexically dot net with your original purchase details and you will be entitled to a
50% discount on the current version.

13.78 **registration is not correct**

*Registration is not correct*

It doesn't match up with what's required for a full updated version! The old registration code in earlier versions is no longer in use. WordSmith will still run but in [Demonstration Version](#) mode.

13.79 **short of memory**

*Short of Memory!*

An operation could not be completed because of shortage of [RAM](#).

13.80 **source folder file(s) not found**

*Source Folder file(s) not found*

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the filename itself.

13.81 **stop list file not found**

*Stop list file not found*

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.82 **stop list file not read**

*Stop list file not read*

Something has gone wrong with a disk reading operation. The file you're trying to read in may be corrupted. This happens easily if you often handle very large files, especially if it's a long time since you last ran *Scandisk* to check whether any clusters in your files have got lost. See your DOS or Windows manual for help on fragmentation.

13.83 **tag file not found**

*Tag File not found*

You typed in the name of a non-existent file. If typing in a file name, remember to include the full drive and folder as well as the file name itself.

13.84 **tag file not read**

*Tag list file not read*

Something has gone wrong with a disk reading operation. The file you're trying to read in may be corrupted. This happens easily if you often handle very large files. See your Windows manual for help on fragmentation.
13.85 **this function is not yet ready**

This function is not yet ready!
Temporary message, for functions which are still being tested.

13.86 **this is a demo version**

This is a demo version
You will probably want to upgrade to the full version.

13.87 **this program needs Windows XP or greater**

This program needs Windows XP or better
From version 4.0, this program has required operating systems for this millennium.

13.88 **to stop getting this message** ...

Get an update. This is "annoyware" for the demonstration version.

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 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
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 mode. Download another from the Internet.

13.100 WordSmith version mis-match

Since the various Tools are linked to each other, it is important to ensure that the component files are compatible with each other. If you get this message it is because one or more components is dated differently from the others.

Solution: download those you need from one of the contact websites.

13.101 XX days left

Message for limited period users only. At the end of this time WordSmith will revert to demo mode.
Index

- # -

# in clusters 270
# symbol 304

- . -

.DOC files
   convert to .TXT in MS Word 436
   convert to plain text using Text Converter 364

.PDF to plain text 364
.XLS to txt 364

- { -

{CHR( conversion 354

- ~ -

~ operator 197

- 2 -

25 lines 449

- 3 -

32-bit version 445

- 5 -

500 key words 247

- A -

about option 429
accents 412
accents & symbols 412
accessing previous results 96
accurate sort in WordList 301
acknowledgements 408
add value to corpus 145
adding notes to data 29
adjust settings 29
Administrator rights 455
Adobe .pdf to plain text 364
advanced concordance settings 159
advanced scripting 104
advanced settings
   convert from UTF-8 31
   customising menus 31
   deadkeys 31
   force keyboard 31
   Maori 31
   menu shortcuts 31
   popup menu 31
   advanced settings button 78
alt-tab 123
annotate source texts 145
API 408
   altering your data 65
   custom .dll file 65
   lemmatising with custom .dll 65
apostrophe 121
apostrophes -- curling or straight 360
Apple Mac 432
Application programming interface 408
associate defined 225
associate word-lists and concordances with file-types 420
associated entries
   lemmas 262
associates 226
attach date to text file 47
auto sentence handling 142
auto-joining lemmas 265
automated file-based concordancing 157
automated processing 104

- B -

Babbage 451
batch choosing 228
batch concordancing 161
batch processing
   and Excel 39
bibliography 409
Big5 365
blanking out entries 164
BNC
  handling of sentences and headings 142
  selecting between texts 133
  selecting within texts 134
  tag file 137
  text format 427
BNC Sampler version 419
boolean and/not 197
bugs 409
burstiness 438

- C -
calculating a plot 242
call a concordance 231
calling other tools 430
cannot compare word-lists in different languages 461
can't see Concord tags 448
CD-ROM
  speed 441
  storage 439
change language of existing data 411
changing font 76
changing from edit to type-in mode 420
Character Profiler
  how to profile text 397
  overview 6
  purpose 397
  settings 400
character sets 411
characters and letters
  accents window 28
  wildcards 28
characters for different languages 412
characters in save as text 216
characters within word 121
chargram
  definition 417
chargram procedure 401
chargrams 418
  display 402
  overview 7
  purpose 401
  settings 405
Charles Babbage 451
check current version 24
checking for updates 78
Chinese Big5 365
Chinese GB2312 365
chi-square 238
chm files not visible 23
Choose Languages
  overview 7
choosing files from standard dialogue box 50
choosing reference corpus 236
choosing texts 42
  .DOC files 42
  .DOCX files 42
  .PDF files 42
  .TXT files 42
  clear previous selection 44
  Dickens text 44
  store text files 44
classroom use
  class instructions 50
  setting up a training session 50
  university or school work 50
clipboard 414
clipboard advanced settings 36
clumps
  regrouping 244
cluster
  definition 417
  cluster settings 312
clusters 441
  joining 274
  reduction & merging 274
clusters in KeyWords 239
Cocoa tags 361
codepages 411
codes 411
codes in search-word 155
collocate follow 180
collocate minimum frequency & length 220
collocates 175
  display 177
  follow 180
  highlighting in concordance 182
  horizons 176
  lemmas 179
  minimum frequency 176
convert within text files 354
converter 346
converting BNC XML version 367
converting Treetagger text 362
copy
  all 64
  choices 64
  selective 64
  specify 64
copy data to Word 414
Corpus Corruption Detector
  aim 321
  overview 7
  process 321
corpus paragraph-count 289
corpus sentence-count 289
corpus word-count 289
correcting filenames 108
count data frequencies 64
crash 409
cumulative scores 61
curly quotes 360
custom column headings 69
custom layouts
  removing 111
custom processing 65
cut spaces 216
cutting line starts 134
- D -
data as text file 242
database construction 231
database statistics 236
date format 417
dates of texts 47
default colours 58
defaults 110
  .ini files 110
  network 110
defining multimedia tags 143
definition of associate 225
definition of key key-word 234
definition of sentence 418
definitions
  chargram 417
classer 417
concgram 386
headings 417
paragraphs 417
sentences 417
token 417
type 417
valid character 417
word 417
deleting entries 125
demo limit 449
demonstration version 419
detailed consistency 256
dice coefficient 260
relation statistics 260
details of MSWord text 332
dice coefficient
  formula 425
dice coefficient for detailed consistency 260
Dickens text 438
directories 423
dispersion 187
dispersion plot
  sorting 205
DOS to Windows 365
download new version 24
downloaded text problems 346
drag and drop 419
drop a text file onto WordSmith 419
duplicate concordance lines 203
duplicate text files 340
- E -
edit mode 420
editing
  column headings 69
  concordances 201
  delete if 69
  delete to end 69
  random deletion of entries 68
  reduce data to N entries 68
  restore to end 69
  reverse deletion 69
  WordList entries 70
encrypt your source texts 366
Entitities to characters 361
Error messages

.ini file not found 455
Administrator rights 455
base list error 456
can only save words as ASCII 456
cannot compute KWs 456
can't call other tool 456
can't make folder as that's an existing filename 456
can't merge list with itself! 456
can't read file 456
character set reset to <x> to suit <language> 457
concordance file is faulty 457
concordance stop list file not found 457
confirmation messages - okay to re-read 457
conversion file not found 457
couldn't merge KW databases 469
destination folder not found 458
disk problem -- file not saved 458
dispersions go with concordances 458
drive not valid 458
expiry date 470
failed to access Internet 458
failed to create new folder 458
failed to read file 458
failed to save 459
file access denied 459
file contains "holes" 459
file contains none of the tags specified 459
file not found 459
filenames must differ 460
for use on pc named XXX 460
form incomplete 460
full drive & folder name needed 460
function not working properly yet 460
invalid concordance file 460
invalid file name 461
invalid KeyWords database file 461
invalid KeyWords file 461
invalid WordList comparison file 461
invalid WordList file 461
joining limit reached 461
KeyWords database file is faulty 462
KeyWords file is faulty 462
limit of file-based search-words reached 462
links between Tools disrupted 462
match list 462
must be a number 462
mutual information 463
network registration used elsewhere 463
no access to text file - in use elsewhere? 463
no associates found 463
no clumps identified 463
no clusters found 463
no collocates found 463
no concordance entries found 464
no concordance stop list words 464
no deleted lines to zap 464
no entries in KeyWords database 464
no fonts available 464
no key words found 464
no key words to plot 465
no KeyWords stop list words 465
no lemma list words 465
no match list words 465
no room for computed variable 465
no statistics available 465
no stop list words 465
no such file(s) found 465
no tag list words 466
no word lists selected 466
not a current WordSmith file 466
not a valid number 466
not a WordSmith file 466
nothing activated 466
only x% of words found in reference corpus 467
original text file needed but not found 467
printer needed but not found 467
read-only folder 460
registration string is not correct 468
registration string must be 20 letters long 467
short of memory 468
source folder file(s) not found 468
stop list file not found 468
stop list file not read 468
tag file not found 468
tag file not read 468
the program needs Windows XP or greater 469
this function is not yet ready 469
this is a demo version 469
too many requests to ignore matching clumps 469
too many sentences 469
truncating at xx words 469
### Error messages
- two files needed: 469
- version mis-match: 470
- why did search fail?: 469
- word list file not found: 470
- word list is faulty: 470
- WordList comparison file faulty: 470
- WordSmith already running: 470
- XX days left: 470

### Example
- Example of aligning: 374
- Example of key words: 233

### Excel
- column totals: 100
- convert to .txt: 364

### Exercises
- 164

### Exiting
- 99

### Export
- index data: 277
- export to spreadsheet etc.: 100

### External drive
- folder letters: 423
- external hard drive: 21
- extracting from text files: 351

### File associations
- 420

### File Utilities
- compare 2 files: 339
- dates of texts: 344
- dodgy text: 346
- editing text file dates: 344
- file chunker: 340
- find duplicates: 340
- find holes in texts: 346
- index: 334
- moving files to different folders: 343
- overview: 8
- rename: 341
- Splitter: 8

### File Utilities Joiner
- joining text files: 338

### File Utilities Splitter
- bracket first line: 335
- end of text separator: 335
- end-of-text symbols: 337
- filenames: 336
- index: 335
- purpose: 335
- symbols: 337
- wildcards: 337

### File Viewer
- overview: 8
- file-based lemmatisation: 265
- file-based search-words or phrases: 157
- filename and path: 214
- filenames
  - display: 109
  - editing: 108
  - tab: 109
- file-types: 420
- find files containing words: 261
- find files with KWs: 73
- find which files contain a word or cluster: 261
- finding a word
  - word or part of a word: 107
- finding by typing: 107
- finding relevant files: 73
- finding source texts: 422
- first use of WordSmith: 438
- flash drive folders: 423
- folder letters: 423
- folder settings: 76
- folder view: 78
- folders: 423
  - force detailed view: 78
- folders created using text converter: 350
- follow-up concordancing: 167
- fonts
  - greek: 76
  - russian: 76
- force folders to show in detailed view: 78
- formulae
  - dice coefficient: 425
  - Log Likelihood: 425
  - MI: 425
  - MI3: 425
  - t-score: 425
  - z-score: 425
- freeze columns: 89
- frequencies of suffixes: 294
full lemma processing  177, 247

- G -

GB2312  365
genral settings  78
get favourite text selection  50
getting started  2
getting started with Concord  14
getting started with KeyWords  15
getting started with WordList  17
globaity of plot  438
green marking in left column  440
grow a concordance line  162
grow and shrink  161

- H -

handling hyphens  121
handling multiple windows  123
handling of numbers  121
handling Word .doc files  436
hash representing words with numbers  304
header removing  130
headings
  definition  417
  start & end  142
hex  332
hide tags  196, 216
hide words  216
highlighting collocates in concordance  182
history list  96
holes in file  459
horizons  176
hotkey combinations  431
hotkeys
  Ctrl/F2  99
  list  431
  space bar  164
how many words  429
how much text  429
how to build a database  231
HTML  427
HTML headers
  cutting out  130
HTML/BNC/XML entities to characters  361
hyphen treatment  427
hyphens  121

- I -

idioms  441
ignore punctuation  219
illegible  450
importing text into a word list  297
incompatibility between word lists  237
index lists
  export  277
  making Wordlist Index  268
  uses  267
  viewing  275
index relationships  285
index settings  312
information about WordSmith version  444
installing WordSmith Tools  21
instructions folder  23
interface  428
international versions  428
introduction to WordSmith Tools  2
inverted commas  448
it won't do what I want  448

- J -

Japanese ShiftJis  365
joining clusters  274
joining entries  262

- K -

key key word defined  234
key key-words  236
key word procedure setting  247
key word settings in Controller  247
keyboard  431
keyness
  definition  235
  p value  234
  thinking about  236
key-ness defined  235
keyness scores  236
keyword database
Index

keyword database 237
related clusters 237
statistics 236

KeyWords
advice 237
associates 225
calculation 238
choosing your files 228
clumps 230
clusters 239
compute a word list 242
database 231, 236
disambiguation 230
display 246
example 233
failure/problems 237
finding KWs in other texts 73
index 223
key key-words 234
links 240
overview 6
plot 242
purpose 223
sorting 245
starting tips 15
tips 237
keywords minimal processing 247
KeyWords plot
display 242
Korean Hangul 365
KWs in other text files 73

- L -

language
Baltic 79
Central European 79
Cyrillic 79
Greek 79
Portuguese 79
Russian 79

language settings
characters within a word 120
end of heading marker 120
end of paragraph marker 120
end of sentence marker 120
heading marker 120
headings (specifying) 120
hyphens 120
numbers 120

Languages Chooser
apostrophe 82
font 84
language 82
new language 84
saving settings 85
sort order 84

Languages Chooser: overview 81
layout
alignment 85
column width 85
decimal places 85
editing column headings 85
grid format 85
headings 85
save layout 85
typeface 85

lemma file 266
lemma list 266
lemma visibility settings 308
lemmas 262
auto-joining 265
file 266
file-based 265
joining automatically 265
joining manually 263
matching in WordList 266
template 265
visibility 308

lemmatising source texts 366
lemmatising using a template 265
letter-count 289
licence details 22
limitations 429
links between tools 430
list of menu options 433
localisation 428
log file to trace problems 32
log likelihood 238
formula 425
log likelihood computing 285
Log Likelihood score 280
logging 32
lowest possible value for clusters 270
LY endings in a word list 294
- M -

Mac version 432
machine requirements 432
make a word list from keywords data 242
manual for WordSmith Tools 432
manual joining 263
manual lemmatisation 263
mark_up
  custom settings 129
  custom settings for BNC tags 130
  document header removal 130
  entity references 141
  tags as selectors 130
  types of 141
marking 263
marking context-word in txt 206
marking entries
  green margin 90
  unmarking 90
  white margin 90
marking search-word in txt 206
mark-up 127
  autoload tag file 128
  colours 137
  handling tag-types 128
  HTML & SGML tags 128
  making a tag file 137
  multimedia 143
  section tag 137
  selecting between texts 133
  selecting within texts 134
match list
  filtering 90
  mark words in a word list 90
mean and standard deviation 289
memory stick 21
memory usage 439
menu choices 433
merge concordances 253
merge wordlists 253
MI score 280
MI3
  formula 425
MI3 computing 285
MI3 score 280
Microsoft Word 414, 436
Minimal Pairs
  aim 323
  choosing files 325
  output 328
  overview 8
  requirements 324
  rules and settings 329
  running the program 330
minimal processing (Concord) 220
minimal processing (KeyWords) 248
modify source texts 145
MS Word 436
multimedia tags 143
multiple file analysis 236
multiple lists 39
multi-word unit 145
mutual information
  computing 285
  formula 425
mutual information scores 280
mutual information screen 280
- N -
nag message 469
nearest tag 194
negative keyness 235
negative keywords 247
network settings 23
network speed 441
network version 23
networks
  defaults 110
new in version 6 4
new user 438
n-grams 441
n-grams in WordList 270
no hts within X words 216
non-characters allowed within a word 121
notes 29
number handling 121
number of concordance entries 216
number sort 203
numbering paragraphs
  Viewer & Aligner 379
numbering sentences
  Viewer & Aligner 379

numbers
  how treated 438

numbers in words: display 304

- O -

omit # in clusters 270
online screenshots 4
options for defaults 110
ordering details 419
over-writing 347
Oxford University Press 419

- P -
p value 234
padding out the search-word with space 214
paragraph
  start & end 142
paragraph marker 120
paragraph numbers
  Text Converter 354
paragraphs
  definition 417
  specifying 120
pasting
  as graphic or as text 414
  concordance into Word 414
  paste special 414
patterns
  highlighting in concordance 182
pen drive 21
percentages v. raw numbers 215
permanent settings 110
phrase frames 272
phrases 441
plain text 42
plot dispersion calculation 438
plot dispersion value 438
plot with grouped files 190
plots and links 240
plotting key words 242
precomposed characters 36
prefix frequencies 294
previous lists 96
price 419
print preview
  zoom 96
printer settings 78
printing
  blank print page 96
  footer 96
  header 96
  landscape 96
  portrait 96
programming WordSmith 408
prompts to save 36
punctuation breaks 184
purple marks in word list display 304

- Q -

quitting 99
quotation marks 448

- R -

RAM availability 439
randomised concordance entries 216
range 255, 256
raw numbers 214, 215
raw numbers v. percentages 215
re-compute filenames after zapping 125
recompute token count 288, 312
recomputing plot 190
reference corpus 439
registry 420
regrouping clumps 244
relationship between collocate and search-word 176
relationship computing
  Log Likelihood 280
  MI statistic 280
  MI3 statistic 280
  T score 280
  Z score 280
relationships computed from an index 285
relationships screen 280
relationships: case sensitivity 177
removable drive 21
remove all mark-up from a corpus 361
remove custom layouts 111
remove duplicates 203
remove line-breaks 358
remove messages 111
remove some XML mark-up 363
rename numerous files 341
re-ordering 125
re-ordering word lists 70
repeated concordance lines 203
replacing 347
report on a crash 409
Requirements 4
re-sorting
  collocates 185
restore all defaults 110
restore factory defaults 36
restore factory settings 110
restore last file 439
restore last work 78
restore settings 111
restricted search 197
ruler 242
running words 417

- S -

save as
  Excel 100
  HTML 100
text 100
XML 100
save favourite text file set 50
save prompts 36
saving
  Ctrl/F2 99
  partial 99
saving defaults 110
scripts 104
search & replace 108
search by typing 420
searching 107
searching by typing 107
search-word
  advanced functions 159
  alternative search words 155
  ascii codes for searching 155
  asterisk 155
  boolean or 155
  case sensitivity 155
  CHR 155
  file-based 157
  history list 155
  slash 155
  syntax 155
  text file specifying 157
  whole word 155
search-word marking in txt file 206
section
  start & end 142
selecting multiple entries 440
selecting within texts 134
sentence
  auto handling 142
  start & end 142
sentence breaks 184
sentence lengths exporting 277
sentence marker 120
sentence only 216
sentences
  definition 417
  specifying 120
separate search-words 216
Set column 164
set column colours 51
set textual date 47
settings 110
  advanced 78
colours 58
defaults 110
folders 76
fonts 76
general 78
language 79, 120
main 78
permanent 110
printer 78
restoring 111
show help 78
statusbar 78
toolbar 78
ShiftJis 365
shortcuts 431
show help at startup 110
show help file 78
show or hide data below a minimum threshold 85
show or hide tags 196
shrink a concordance line 162
SI numbers 121
single words 441
slow 451
sorting
  Concord 203
  consistency lists 289
  dispersion plot 205
  KeyWords 245
  tags 194
  word list 301
sound & video tagged files 207
sound file tags 143
source text(s)
  view 112
source texts 422
  converting to a better format 358
  modify 145
speed 441
standardised or mean type/token ratio 299
start and end of sentence 142
statistics
  headings 289
  letters 289
  paragraphs 289
  sections 289
  sentences 289
  word list statistics 289
  word-count 289
statistics: vertical or horizontal 291
status bar 433, 442
statusbar 78
stop at punctuation 184
stop at sentence break 184
stop lists 116
stop lists v. match lists 296
stoplist.cod 355
stopping 119
storage 439
suffix frequencies 294
summary statistics (general) 64
suspending processing 119
swap tags and words 361
symbols 412

- T -

  T score 280
tag concordancing 193
tag file 137
tag types 141
tag visibility 196
tag-free corpus 361
tagged text 127
tags
  overview 127
tags in WordList 302
tags swapped with words 361
tags to exclude 137
tags to retain 137
teacher instructions 50
Test for Unicode 46
text characteristics 120
Text Converter 358
  .DOC 364
  .DOCX 364
  .Excel 364
  .PDF 364
  asterisk 356
  BNC XML version conversion 367
collection file 355
cutting header 347
  extracting 351
  folders 347
  index 347
  insert numbering 354
  into Unicode 358
  just one change 354
  line-breaks removal 358
  move if 352
  numbers: insert paragraph numbers in your corpus 354
  overview 9
  purpose 346
  removing all tags 356
  sample conversion file 357
  settings 347
  syntax 356
  Unicode conversion 358
  UTF16 conversion 358
Text Converter 358
  UTF8 conversion 358
  wildcards 356
text date analysis 122
text file
  use to build a word list 297
text file dates 344
text formats 120
text segments in Concord 212
texts 44
  choosing 44
  favourites 50
  more texts 44
tie-breaking 203
time-lines 122
to right only 286
token
  definition 417
token count 288
token recomputing 312
toolbar 78, 433
tools for pattern-spotting 443
tool-specific limitations
  Concord 429
  KeyWords 429
  Text Converter 429
  Viewer & Aligner 429
  WordList 429
training students 50
Treetagger conversion 362
Troubleshooting 448, 451
  accented symbols 449
  apostrophes not found 448
  colours unreadable 450
  column spacing 448
  Concord tags problem 448
  Concord/WordList mismatch 449
  crashed 449
  curly quotation marks 448
  demo limit 449
  keys don't respond 450
  pineapple-slicing 451
  potato-peeling machine 451
  printer won't print 451
  quotation marks not found 448
  smart quotations 448
  takes ages 451
  the English 451
  weird symbols 449
  won't start 451
  WordList out of order 452
t-score
  formula 425
  T-score computing 285
  Two word-list analysis 224
type
  definition 417
type/token ratios 299
type-in mode 420
type-in search 107
types of tag 141
typing characters into Concord 412
- U -
  undefined tags 216
  underscore 121
  underscore tags 361
  Unicode 411
  unicode explained 422
  Unix to Windows 366
  unjoin all entries 263
  unjoining entries 263
  unmarking 263
  unreadable 450
  updater.exe 21
  updating WordSmith 78
  updating your version 21
  USB drive 21
  USB drive folders 423
  user licence 22
  user-defined categories 164
  saving 145
  user-defined processes 65
  UTF8 versus UTF16 422
- V -
  valid character
    definition 417
  value-added annotation 145
  version 4 differences 445
  Version Checker
Version Checker
  overview 9
version checking 24
version date 444
version francaise 428
vertical view of statistics 291
Viewer & Aligner 374, 382
  adjusting with mouse 377
  aligning 374
  aligning -- an example 374
  aligning the sentences 377
  colours 379
dual-text aligning 374
  editing 377
  index 373
Korean and English aligned text 374
  languages 378
mowing sentences 377
overview 11
paragraph numbering 379
purpose 372
reading in your plain text 380
sentence joining 382
sentence numbering 379
settings 383
splitting 382
technical aspects 383
translation mis-matches 383
troubleshooting 384
unusual sentences 384
  viewing options 379
  viewing original text file 161
  viewing the original text 112

- W -

WebGetter
  display 318
  limitations 320
overview 12
purpose 315
settings 317
what is a concordance 155
What's new 4
whole word search 155
why won't it... 448
window management 123
windows
  managing 123
Windows file associations 420
Windows XP 432
word
  definition 417
word cloud settings
    shape 58
word clouds
  collocates 185
  example 124
word count in MS Word 293
Word documents 436
word patterns 202
word separators 419
Word to .txt 364
word_TAG to <TAG>word 361
WordList
  case sensitivity 301
  clusters 270
  coloured tags 302
  comparing word lists 252
  comparison display 254
  compute keywords 256
  create using text file 297
detailed consistency 256
displaying comparisons 254
find files containing a word 261
finding entries 279
  index 251
keys for searching 279
  locating entry-types 279
merging 253
minimum & maximum settings 301
n-grams 270
overview 6
prefix for tag 302
purpose 251
searching using menu 279
  simple consistency 255
sort order 301
  sorting problems 452
starting tips 17
summary statistics 294
tags 302
  tags as prefix 302
the basic display 304
WordList Index
  clusters 312
  computing clusters 441
  n-grams 441
  omit # 270
  relationship settings 312

WordList: altering entries 70
Word’s results are different 293

WordSmith controller
  Concord settings 216
  index settings 312
  KeyWords settings 247
  WordList settings 308

wordsmith exe file (controller) 4

WordSmith Group 417

WordSmith Tools
  installation 21
  manual 432
  version 444

wordsmith6.ini and networks 23

WSConcGram
  aims 385
  definition of concgram 386
  display 389
  exporting concgrams 396
  filtering 394
  generating concgrams 387
  overview 12
  settings 387
  viewing 389

- X -

X-letter word count 289
XML 427
  attributes 148
  entities 148
  parsing XML 148
  text handling 148
XML simplification 363

- Y -

Yasumasa Someya 266

- Z -

Z score 280
zapping
  filenames recomputed after 125
zip files 446
z-score
  formula 425
Z-score computing 285
WordSmith Tools Manual