WordSmith Tools Manual

version 3.0

http://www.liv.ac.uk/~ms2928/wordsmit.htm

Contents

WordSmith Tools

Getting Started

What do the WordSmith Tools do?

Trouble-shooting

What to do if it doesn't do what I want...

Controller

Concord

WordList

KeyWords

Splitter

Text Converter

Viewer

Settings & Defaults

- Buttons
- Choosing Text Files
- Colours
- Defaults
- Directories
- Exiting
- Fonts
- Keyboard Shortcuts
- Layout
- Printer Settings
- Stop Lists
- Tagged Texts
- Text Characteristics

Other Information

- Bibliography
- Character Sets
- Clusters
- Contact Details
- Definitions
- Demonstration Version
- Installation
- Limitations
- Links between the Tools
- Manual
- The Role of Tools
- Version Information
- Window Management

For information on how to use the Windows Help system itself, press F1.

General Stuff

Introduction

Machine Requirements

To run **WordSmith Tools** you need:

- at least 4MB of RAM (8 in Windows 95)
- at least 5MB of hard disk space
- an IBM-compatible PC with a 386 or better processor.
- Windows™ 3.1, 3.11, 95 or NT, 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 .

Installation of WordSmith Tools

You will need about 5 Mb of space on your hard disk for the programs, but during installation you will need double that. (Anyway Windows won't run well without at *least* 10Mb spare.)

1 You have received or downloaded one or more **.exe** files. Put them in a temporary directory, e.g. **c:\temp**. It's easiest if this is a clean directory without any other files in it.

2 Run them. This will expand all the files needed for **WordSmith Tools**. If you're short of disk space, you can delete the **.exe** files now; if not, clean up later.

3 Now run **setup.exe**, which you'll find in the same directory.

4 Choose where you want the definitive version of **WordSmith Tools** to go. I suggest **c:\wsmith** but change it to **d:\wsmith** or **e:\apps\wsmith** or whatever you prefer. **Setup.exe** will copy everything over, though you can choose not to install certain sections of the suite if you like. These choices are visible within **setup.exe**. You can install an icon if you like.

5 Finally, run **c:\wsmith\wshell.exe** to get started. You will need to "update from demo" if you have the registration key; otherwise **WordSmith** will go through its paces as a Demonstration Version.

6 Afterwards, you can delete all the stuff in your **c:\temp** directory.

You can get a more recent version at **http://www.liv.ac.uk/~ms2928/wordsmit.htm**.

To un-install, just delete all the files in your **\wsmith** directory.

See also: Contact Addresses.

Introduction to WordSmith Tools

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

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

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

Getting Help

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.

Press 💡 to get advice, a map (*Where am I?*) and suggestions. This will show you what you might usefully do next: in the case of a word processor you already know what texts are like, but **WordSmith Tools** offer possibilities you may not have considered before. Also, you will want to know what the features are called -- this is a good way to find out and get access to the right part of the help system.

Click here to get started straight away with WordList, Concord, or KeyWords.

WordList: Quick Start

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, on the right hand side of your screen.

the Start button 🟢

Pressing 🟢 takes you to a dialogue box which lets you choose your texts or change your choice, and make a new word list.

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 at the bottom right corner of your screen shows a bar indicating how it's getting on. After **WordList** has finished making the list, you will see three windows showing the words from your text file in alphabetical order and in frequency order, and statistics.

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

See also WordList Help Contents.

Concord: Quick Start

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 *Concord*. Click and **Concord** will open up, on the right hand side of your screen.

the Start button ●

Pressing ● takes you to a dialogue box which lets you choose your texts or change your choice, and make a new concordance.

You will need to specify a Search-Word or phrase.

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

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

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

See also Concord Help Contents.

KeyWords: Quick Start

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, on the right hand side of your screen.

the Start button ●

Pressing ● takes you to a dialogue box which lets you choose your wordlists. You'll need to choose two word lists to make a key words list from: one based on a single text, and another one based on several texts, enough to make up a good reference corpus for comparison.

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

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

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

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

Concord Tool

Concord is a program which makes a concordance using DOS, Text Only, ASCII or ANSI text files.

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

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

See also Concord Help Contents Page, The buttons

WordList Tool

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

See also WordList Help Contents Page, The buttons

KeyWords Tool

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

Text Converter Tool

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 directory 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 up to 16,368 text files, one after the other. As it does so, it can also replace up to 500 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, The buttons

Splitter Tool

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

See also Splitter Help Contents Page, The buttons

Viewer Tool

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

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

See also Viewer Help Contents Page, The buttons

Wshell.exe: the 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.

Manual for WordSmith Tools

This help file exists in the form of a manual, which you get when you install. The file, astonishingly called **manual.doc**, is in Microsoft Word™ format. If you have Word, load it up and print. It comes to about 140 pages of A4 and has a table of contents and a fairly detailed index (which I used **WordList** and **KeyWords** to help me create). Most people find paper easier to deal with than help files!

You may find it useful to see screenshots of **WordSmith** in action: check out Contact Addresses.

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 their name states, are not products but tools. You can use them to investigate many kinds of pattern in virtually any texts written in a good range of different languages.

Insight through Transformation

No, this is not a religious claim! The claim I am making is psychological. It is through changing the shape of data, reducing it and then re-casting it in a different format, that the human capacity for noticing patterns comes to the fore. The computer cannot "notice" at all (if you input 2 into a calculator and then keep asking it to double it, it will not notice what you're up to and begin to do it automatically!). Human beings are good at noticing, and particularly good at noticing visual patterns.

By transforming a text into a list, or by plotting keywords in terms of where they crop up in their source texts, the human user will tend to see a pattern. Indeed we cannot help it. Sometimes we see patterns where none was intended (e.g. in a cloud). There can be no guarantee that the pattern is "really there": it's all in the mind of the beholder.

WordSmith Tools are intended to help this process of pattern-spotting, which leads to insight. The tools in this kit are intended therefore to help you gain your own insights on your own data from your own texts.

Types of Tool

All tools take up positions on two scales: the scale of specialisation and the scale of permanence.

general-purpose ---------------- specialised

general-purpose

The spade is a digging tool which makes cutting and lifting soil easier than it otherwise would be. But it can also be used for shovelling sand or clearing snow. A sewing machine can be used to make curtains or handkerchiefs. A word-processor is general-purpose.

specialised

A thimble is dedicated to the purpose of protecting the fingers when sewing and is rarely used for anything else. An overlock device is dedicated to sewing button-holes and hems: it's better at that job than a sewing machine but its applications are specialised. A spell-checker within a word-processor is fairly specialised.

temporary ---------------- permanent

temporary

The branch a gorilla uses to pull down fruit is a temporary tool. After use it reverts to being a spare piece of tree. A plank used as a tool for smoothing concrete is similar. It doesn't get labelled as a tool though it is used as one. This kind of makeshift tool is called "quebra-galho", literally branch-breaker, in Brazilian Portuguese.

permanent

A chisel is manufactured, catalogued and sold as a permanent tool. It has a formal label in our vocabulary. Once bought, it takes up storage room and needs to be kept in good condition.

The **WordSmith Tools** in this kit originated from temporary tools and have become permanent.

They are intended to be general-purpose tools: this is the Swiss Army knife for lexis. They won't cut your fingers but you do need to know how to use them.

see also : Acknowledgements

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

The first published version was written in Borland™ Pascal with the time-critical sections in Assembler. Subsequently the programs were converted to Delphi™ 16-bit; they are now being re-written in Delphi™ 32-bit so that they can run more efficiently in Windows 95 and NT™

I am grateful to generations of students and colleagues at the Department of English, University of Liverpool, and the MA Programme in Applied Linguistics at the Catholic University of São Paulo, for their feedback on aspects of the suite (including bugs!), and suggestions as to features it should have. Researchers from many other countries have also acted as alpha-testers and beta-testers and I thank them for their patience and feedback. I am also grateful to Nell Scott and other members of my family who have always given valuable support, feedback and suggestions.

Mike Scott

Feel free to email me at my contact address with any further ideas for developing **WordSmith Tools**.

Bibliography

Aston, Guy, 1995, "Corpora in Language Pedagogy: matching theory and practice", in G. Cook & B. Seidlhofer (eds.) *Principle & Practice in Applied Linguistics: Studies in honour of H.G. Widdowson*, Oxford: Oxford University Press, 257-70.

Aston, Guy & Burnard, Lou, 1998, *The BNC Handbook*, Edinburgh: Edinburgh University Press.

Clear, Jeremy, 1993, "From Firth Principles: computational tools for the study of collocation" in M. Baker, G. Francis & E. Tognini-Bonelli (eds.), 1993, *Text and Technology: in honour of John Sinclair*, Philadelphia: John Benjamins, 271-92.

Dunning, Ted, 1993, "Accurate Methods for the Statistics of Surprise and Coincidence", *Computational Linguistics*, Vol 19, No. 1, pp. 61-74.

Fillmore, Charles J, & Atkins, B.T.S, 1994, "Starting where the Dictionaries Stop: The Challenge of Corpus Lexicography", in B.T.S. Atkins & A. Zampolli, *Computational Approaches to the Lexicon*, Oxford:Clarendon Press, pp. 349-96.

Murison-Bowie, Simon, 1993, *MicroConcord Manual: an introduction to the practices and principles of concordancing in language teaching*, Oxford: Oxford University Press.

Nakamura, Junsaku, 1993, "Statistical Methods and Large Corpora: a new tool for describing text types" in M. Baker, G. Francis & E. Tognini-Bonelli (eds.), 1993, *Text and Technology: in honour of John Sinclair*, Philadelphia: John Benjamins, 293-312.

Oakes, Michael P. 1998, *Statistics for Corpus Linguistics*, Edinburgh: Edinburgh University Press.

Scott, Mike, 1997, "PC Analysis of Key Words - and Key Key Words", *System*, Vol. 25, No. 2, pp. 233-45.

Sinclair, John M, 1991, *Corpus, Concordance, Collocation*, Oxford: Oxford University Press.

Stubbs, Michael, 1986, "Lexical Density: A Technique and Some Findings", in M. Coulthard (ed.) *Talking About Text: Studies presented to David Brazil on his retirement*, Discourse Analysis Monograph no. 13, Birmingham: English Language Research, Univ. of Birmingham, 27-42.

Stubbs, Michael, 1995, "Corpus Evidence for Norms of Lexical Collocation", in G. Cook & B. Seidlhofer (eds.) *Principle & Practice in Applied Linguistics: Studies in honour of H.G. Widdowson*, Oxford: Oxford University Press, 245-56.

Tuldava, J. 1995, *Methods in Quantitative Linguistics*, Trier: WVT Wissenschaftlicher Verlag Trier.

Youlmans, Gilbert, 1991, "A New Tool for Discourse Analysis: the vocabulary-management profile", *Language*, V. 67, No. 4, pp. 763-89.

Contact Addresses

Downloads

You can get a more recent version at **http://www.liv.ac.uk/~ms2928/wordsmit.htm** or **http://www.ndirect.co.uk/~lexical/wordsmit.htm**. There are also some free extra downloads (programs, word lists, etc.) there too, with a more complete stock at the latter address. And links to sources of free text corpora.

The latest official Oxford University Press version is at **http://www.oup.co.uk/elt/catalogu/multimed/4589846/4589846.html**. This is usually not as up to date.

Screenshots

visit **http://www.ndirect.co.uk/~lexical/wsmhomep.htm** or **http://www.liv.ac.uk/~ms2928/wsmhomep.htm** 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

email caldwelj@oup.co.uk for current purchase prices.

Complaints & Suggestions

If you do not have the official OUP version but one from my website, please do **not** email OUP but me (Mike.Scott@liv.ac.uk). Please give me as full a description of the problem you need to tackle as you can, and details of the equipment too. Please don't include any attachments over 200K in size. I do try to help but cannot promise to…

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 wordlist, or concordance lines, etc.) is limited to a small number of lines which can be shown or printed. (If you save data, all of it will be saved; it's just that you can't see it all in the demo version.)

Updating

To update your demo version,

1. Send Oxford University Press

a. the current purchase price

b. the *exact* name which you want to use as your registered name for **WordSmith Tools**. This name (at least 8 letters long) appears in the main window and whenever you access the *About* menu option (F9). Be careful to spell this name clearly and *exactly* as you typed them in installation -- to a computer, "Mary J. Smith" is not the same as "Mary J Smith"!

Full details of price and addresses for contacting Oxford University Press are in the OupOrder.txt file in your \wsmith directory.

When these are received, Oxford University Press will contact you, giving you a registration code.

2. Upon receipt of the registration code, run **WordSmith Tools** as usual. You'll see a menu option called *Update from Demo*. Choose this and type in your registration name plus "any other details" such as your place of work, and a 20-letter registration code. Your software will then be fully enabled, and the *Update from Demo* menu option will disappear.

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

BNC Sampler Version

This is a special version which is not restricted in what it'll show. However, it is restricted to running from the CD-ROM and choosing only from the text files there; if any others are chosen you'll be in demo mode.

See also: Version Information, Contact Addresses.

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 only have one copy of each Tool running at any one time. Most of the Tools allow you to work on lots of sets of data at once.

You can process up to 16,368 text files in up to 16,368 directories at any one time.

You can handle up to 2,000 tags in a tag file. Tags to ignore or ones containing an asterisk can span up to 1,000 characters.

tip

Press F9 to see the "About" box -- it shows the version date and how much memory you have available. If you have too little memory left, try a) closing down some applications, b) closing WordSmithTools and re-entering.

See also: Specific Limitations of each Tool

Specific limitations

Concord limitations

You can compute 16,368 lines of concordance using **Concord**.

Concord allows 80 characters for your search-word or phrase, though you can specify up to 300 concordance search-words in a search-word file.

Each concordance can store up to 16,368 collocates with a maximum horizon of 25 words to left and right of your search-word.

Nearest Tags can remember up to 2,000 instances of tags in each 30,000 characters of text processed.

You can find up to 16,368 clusters with the same maximum horizon of 25 words to left and right of the search-word.

WordList limitations

Individual word lists can contain over 8 million separate entries. (To get 8 million separate word types you would have had to read in *many many* thousands of millions of words and these would have to contain a lot of unusual proper nouns. A very large dictionary might contain half a million entries. Word clusters do bump up the number of entries but they are also greedy of machine space.) Each of these words can have appeared in your texts over 2 thousand million times.

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

Individual key word lists can contain 16,368 separate key words.

A key words database can contain data from 16,368 separate text files.

One key-word plot per key-word display. (If you want more, call up the same file in a new display window.)

number of link-windows per key-word plot display: 20.

number of windows of associates per key key-word display: 20.

Splitter limitations

Each line of a large text file can be up to 10,000 characters in length. That is, there must be an <Enter> from time to time!

Text Converter limitations

You can convert up to 16,368 separate files in one batch.

There can be up to 500 strings to search-and-replace for each.

Each search-string and each replace-string can be up to 80 characters long.

An asterisk must not be the first or last character of the search-string.

When the asterisk is used to retain information, the limit is 1,000 characters.

Text Viewer limitations

If you press the View button 📋 when choosing texts, **Viewer** will call up the first 9 source text files selected.

Each one can have up to 16,368 sentences in it. Each sentence can be up to 10,000 characters long (enough for about 1,600 words).

When choosing texts or jumping into the middle of a text (e.g. after pressing the View button 📋 in Concord), **Viewer** will only process 30,000 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 to the full 16,368 sentence limit.

You can only work on one dual-aligned text at a time.

See also: General Limitations

Links between the Tools

The programs in **WordSmith Tools** are linked to each other via wshell.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, directories, fonts, stop lists, etc.

In general, if you press Control-C in **WordList** or **KeyWords** you'll go straight to a concordance, computed using the current word and using the current files.

Press Control-W in **Concord** or **KeyWords** to start a wordlist using the current files.

Each Tool will send as much relevant information as possible to the Tool being called. This will include: the current word (the one highlighted in the scrolling window) and the text files where any current information came from.

Example: after computing a word list based on 3 business texts, you discover that the word *hopeful* is more frequent than you had expected. You want to do a concordance on that word, using the same texts. Place the highlight on *hopeful,* hold down Control and press C. Now you can see whether *hopeful* is part of a 3-word cluster, or view a dispersion plot.

Example: after computing a key words database using 300 business texts, you discover that the word *bid* seems to be a key key-word, and that it's associated with *company*, *shares* etc. Place the highlight on *bid,* press Control-C and a concordance will be computed using the same 300 texts. Now you can check out the contexts: is *bid* a bid for power, or is it part of a tendering process?

Example: you have a concordance of *green*. Now press Control-W to generate a word list of the same text files. Press Control-K to compare this word list with a reference corpus list to see what the key words are in these text files.

RAM and Storage

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). In Windows 3.x, the first line concerning memory tells you how much total memory there is, including use of virtual memory management with a hard disk swap file, followed by the amount without virtual memory management. In Windows 95 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 8MB of RAM or more you should be able to have a copy of a wordlist based on millions of words of text, and at the same time have a powerful word-processor and a text file in memory.

See also: speed

Speed

To make a wordlist on 4.2 million words used to take about 20 minutes on a 1993 vintage 486-33 with 8Mb of RAM. The sorting procedure at the end of the processing took about 30 seconds. A 200Mz Pentium with 64MB of RAM handled over 1.7 million words per minute. On a 100Mz Pentium with 32Mb of RAM this whole process took about 3 and a half minutes, working at over a million words a minute.

When concordancing, tests on the same Pentium 100, using one 55MB text file of 9.3 million words, and a quad-speed CD-ROM drive, showed

search-word	**source**	**speed**

quickly	CD-ROM	6 million words per minute
quickly	hard disk	12 million wpm
the	CD-ROM	900,000 wpm
the	hard disk	1 million wpm
thez	CD-ROM	6 million wpm
thez	hard disk	16 million wpm

Tests using a set of text files ranging from 20K down to 4K, using *quickly* as the search-word, gave speeds of 2 million wpm rising with the longer files to 4 million wpm. Making a word list on the same set of files gave an average speed of 800,000 wpm. On the 55MB text file the speed was around 1.35 million wpm.

These data suggest that factors which slow concordancing down are, in order, word rarity (*the* was much slower than *quickly* or the non-existent *thez*), text file size (very small files of only 500 words or so (3K) will be processed about three times as slowly as big ones) and disk speed (the outdated quad speed CD-ROM being roughly half the speed of the 12ms hard disk). When Concord finds a word it has to store the concordance line and collocates and show it (so that you can decide to suspend any further processing if you don't like the results or have enough already). This is a major factor slowing down the processing. Second, reading a file calls on the computer's file management system, which is quite slow in loading it, in comparison with Concord actually searching through it. Third, disk speeds are quite varied, floppy disks being much the worst for speed.

If processing seems excessively slow, close down as many programs as possible and run **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.

Entomology

All computer programs contain bugs. You may have seen a "General Protection Fault" message when using big expensive drawing or word-processing packages. If you get one of these in any of the programs, you can usually 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.

error messages

These warn you about problems which occur as the program works, e.g. if there's no room left on your disk, or you type in an impossible filename or a number containing a comma.

See also: troubleshooting.

Version Information

This help file is for version 2 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. Your registration code can be used on updated versions.

See also: Demonstration Version, Contact Addresses.

Adjustments

The Buttons

Most button functions apply to the current window of data -- the one whose caption bar is highlighted.

Where applicable, one or more of the following buttons will be available at the top of each display window. You can also right-click in the window to obtain a menu of the same functions.

advice

opens a window showing a map of **WordSmith Tools**, giving a view of where you are now and where you might go next; also offers advice depending on the Tool.

associates

opens a new window showing Associates.

auto-join

joins (lemmatises) automatically.

auto-size

re-sizes each line of a display so that each one shows as much data as it should. Most windows have lines of a fixed size but some, e.g. in **Viewer**, allow you to adjust row heights. This adjusts line heights according to the current highlighted column of data.

change case

changes case (*lower*, *First Letter*, and *CAPITALS*).

clumps

computes clumps in a keywords database

regroup clumps

regroups the clumps

clusters

computes concordance clusters.

collocates

shows collocates using concordance data.

compute

calculates a new column of data based on calculator functions and/or existing data.

redo collocates

recalculates collocates, e.g. after you've deleted concordance lines.

concord

within **KeyWords, WordList**, starts **Concord** and concordances the highlighted word(s) using the original source text(s).

copy

allows you to copy your data to a variety of different places (the printer, a text file, the clipboard, etc.).

double columns

allows you to double the number of columns, so as to save paper when printing.

edit

allows editing of a list **or** searches for a word (type-in search).

filenames

opens a new window showing the filenames from which the current data derived. If necessary you can edit them.

find files

finds any text files which contain *all* the words you've marked.

grow

increases the height of all rows to a fixed size. See shrink (▬) below.

help (also F1)

opens WordSmith Help (this file) with context-sensitive help.

join

joins one entry to another e.g. sentences in **Viewer**, words in **WordList** (lemmatisation).

keep notes

allows you to add notes and save them with your data.

layout

This allows you to alter many settings for the layout: the colour of each column, whether to hide a column of data, typefaces and column widths.

links

computes links between words in a key-words plot.

mark

marks an entry for joining or finding files.

match lemmas

checks each item in the list against ones from a text file of lemmatised forms and joins any that match.

match list

matches up the entries in the current list against ones in a "match list file" or template, marking any found with (~).

mutual information

computes mutual information scores in a WordList index list.

patterns

computes collocation patterns.

plot

opens a new window showing a Concord dispersion plot or KeyWords plot.

plot ruler

shows/hides text divisions in a KeyWords plot.

print (also F3)

previews your window data for printing; can print to file, which is equivalent to "save as text".

refresh

re-reads your text file (in Viewer) or re-draws the screen (in Print Preview).

replace

search & replace, e.g. to replace drive or directory 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.

save (also F2)

saves your data using existing file-name; if it's a new file asks for file-name first.

save as

saves after asking you for a file-name.

save as text

saves as a **.txt** file: plain text. This is ***not*** the same as saving your current data so as to get it back later within **WordSmith Tools**.

search (also F12)

searches within a list.

shrink

reduces the height of all rows to a smaller fixed height. See grow (🟪) above.

skim

in **Viewer**, allows timed skimming through a text.

slide

slides all the buttons to the left (the leftmost wraps to the right). This is to help you see the buttons whenever not all of them would otherwise be visible.

start

gets you started in the various Tools, e.g. to make a concordance, a word list, or a key words list.

statistics

opens a new window showing detailed statistics.

summary statistics

opens a new window showing summary statistics, e.g. proportion of lemmas to word-types.

swap columns for rows

swaps the columns and rows. **WordList** statistics are shown by default with the file data in each column. Click this button to swap the row data with the column data.

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.

wordlist

within **KeyWords**, makes a word list using the current data.

zap

zaps any deleted entries.

see also: Keyboard Shortcuts

Keyboard Shortcuts

scrolling windows:

Control-Home	to top of scrollable list
Control-End	to last line of list

if it's ordered alphabetically, type-in your search-word

and if it scrolls horizontally:

Home	to left edge
End	to right edge
Control-Right	one word to right
Control-Left	one word to left

hotkeys:

Ctrl-C	call **Concord** from within another Tool
Ctrl-W	call **WordList** from within another Tool
Ctrl-Ins	copy blocked section to clipboard

Shift-cursor keys	block a section
F1	help
F2	save results
F3	print results
F4	join entries
F5	mark entries for joining
F6	re-sort
F7	view source text
F8	seek short sentences (**Viewer**)
F9	About box (shows version-date and memory availability)
F12	search within a list
Ctrl-A	open a word list (**A**lphabetical)
Ctrl-B	make a **B**atch of files (WordList and KeyWords)
Ctrl-D	make a KeyWords **D**atabase
Ctrl-F	open a word list (**F**requency)
Ctrl-M	**M**erge 2 word lists or KeyWords databases
Ctrl-R	**R**etrieve a word list (alphabetical window + frequency window)
Ctrl-S	open a word list (**S**tatistics)
Alt-H	access to ***H**elp* sub-menus
Alt-S	access to ***S**ettings* sub-menus
Alt-W	access to ***W**indow* sub-menus
Alt-X	e**X**it the tool
Alt-Z	Zap deleted lines

see also: Buttons

Batch File Names

Just before processing a batch of files, you'll get a dialogue box showing how many files you've selected, and suggesting a directory for the lists to be stored in. You can change this directory if you like: if the directory you choose doesn't exist, it will be created if the name you type is legal.

The file names can be based either on the original names or on a mask.

Original Names

Names are based on the original file names. That is, an existing file name has a number added to it or its last letter(s) changed to a number so as to guarantee that it is unique in that directory. The number starts at 1 and goes up to the number of files being created in the batch.

Mask

This option allows you to specify a name, up to 6 characters in length, which will serve as a basis for the new word lists you're creating. If you type in **maria**, for example, the corresponding word list file will be **maria001.lst**, and subsequent ones will be **maria002.lst**, **maria003.lst**, etc. (A key word list batch would, by default, have the ending **.kws.**) **Obs** would generate **obs00001.lst**, **obs00002.lst**, **obs00003.lst**, etc.

store in a database

The batch of word lists or key word lists can be stored as separate files, or stored all within one big computer file (a database). The advantage of this database strategy is that it wastes much less hard disk space; the disadvantage is that it's trickier to access the individual lists within the big file.

See also: KeyWords Batch Processing, Wordlist Batch Processing

Character Sets

DOS

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

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

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

Windows

In Windows the codes are different again. (The **£** symbol is code 156 in DOS but 163 in Windows.) For Greek, Central European or Baltic languages, Turkish or Cyrillic you can choose special Windows code-pages; you will need to enable the appropriate fonts for Windows 3.1 to see them properly; in Windows 95 you can get such non-Western fonts enabled via Microsoft Plus. If your texts were written using a Windows word-processor and saved as text in Windows, the accented characters will obey the Windows codes. You will have access to a few more symbols than in DOS (e.g. **®,©,™** and curly apostrophes).

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

Tip

To check results after changing the code-page, select Choose Texts and *View* the file in question. While viewing you can change *Text Characteristics* until it looks right. If you can't get it to look right, you've probably not got a cleaned-up plain text file but one straight from a word-processor. In that case, take it back into the word-processor and save it as text again as a plain text file in Windows format, which is more up-to-date than DOS formats.

see also: Choosing Accents & Symbols, Windows Symbols; Accented characters; Choosing

Language

Choosing the Language

Choose the language for the text you're analysing in the Controller under *Adjust Settings | Text*. The language and format must be compatible, e.g. English is compatible with Windows Western (1252), DOS Multilingual (850). See below for a table.

WordSmith Tools handles a good range of mostly European languages, ranging from Albanian to Ukrainian and up to 12 others which you can define for yourself. Unfortunately Chinese, Japanese, Arabic etc. are not handled, unless you choose to transliterate them using an alphabet from another language such as English or Russian. You can view wordlists, concordances, etc. in different languages at the same time.

The point of it…

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

The alphabet information is stored in a file called **langinfo.txt**, which you'll find in your WordSmith directory along with the programs. This contains definitions for the languages; you may alter them if you wish to handle a new language; or else contact me and ask me to include the language name and details in the next version. The definition information includes: alphabetical order, handling of accents, cases where two or three symbols are treated as one, or where one is considered equivalent to two others (**ß** and **ss** in German, for example). Full details are supplied in the header to **langinfo.txt**.

Usually, you will specify only upper case letters in **langinfo.txt**, but if you need to distinguish between upper and lower case letters in a word list (so *the* is counted separately from *The* and *THE*) you can specify your own preferred sort system as an Other language, and activate case sensitivity in **WordList**. Warning: if you choose to define your own language features, keep a backup safe. The alphabet defined in it will always be needed to view any lists you save when it is in use.

Many languages have outdated special DOS formats only of use with "legacy text".

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

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

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

Estonian, Latvian, Lithuanian

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

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

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

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

Windows **Greek** (1253) handles Greek

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

There are 12 "Other Languages" which you can define for yourself in the **langinfo.txt** file.

See also: Choosing Accents & Symbols, Windows Symbols; Accented characters

Choose Text(s)

This function is accessed from the File menu in the Controller and the Settings menu or Start button (●) in the various Tools

The list box shows full file details (name, date, size). In versions of Windows prior to 95 it can hold up to about 4,700 entries, though if there are more than 1,200 you'll see partial details only. Windows 95 does not have this limitation.

See choosing more than one file for tips on selecting more than one text file at once..

The default file specification is ***.*** (i.e. all files) but this can be altered in the bottom left corner or set permanently in wshell.ini.

Sorting

By clicking on N,T, D or S you can re-sort the directory listing by *Name*, *Type, Date* or *Size*. The default sort is by alphabetical *Name* order, unless there are lots of files, in which case the default is by size.

Drives and Directories

There's a listbox for changing drives. Clicking on a yellow blob in the directory listing opens up any sub-directories.

View

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

Favourites

Allows you to save or get a previous file selection as a plain text file, saving you the trouble of making and remembering a complex set of choices.

All

All selects all the files in the current directory. If you choose this, you will get a chance to choose all the files in all sub-directories too.

Store, OK

Store puts the current file selection into store, ready for you to change directory and choose some more. Press OK if you've finished selecting altogether.

You can check on which ones have been selected under *All Current Settings*.

The limit for numbers of files and directories is 16,368 (of each). Duplicates will be filtered out unless they're in different directories.

Clear Previous Selection

This button is visible if you have a selection stored already. As its name suggests, this allows you to change your mind and start afresh.

Tip

If you want to choose a mixed set of text files and their names reflect their contents, you could sort them by size or date instead of by name in the *Choose Text Files* menu item.

Choosing files from standard dialogue box

For some procedures, you may wish to choose more than one file using a standard Windows dialogue box (the kind which allows you to choose fonts, set up your printer, or in this case open files).

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.

Favourite Texts

Clicking on *Favourites* opens a dialogue box which allows you to save your currently selection of texts or read a previously-saved selection from disk. This is useful if you often want to select the same set of files, stored in various different directories. By default the filename will be **wshell.tmp**, in your main **WordSmith** directory. You may edit the resulting file using **notepad**, but note that each file needed must be fully specified: wildcards are not used and a full drive:\directory path is needed.

If you choose to select within texts, looking for certain tags and using them as selectors, you will get a chance to save the names of all files which met your criteria, in a file called **favour.txt**, in your main **WordSmith** directory.

See also: Choosing Texts

Single Words v. Clusters

The point of it…

Clusters are words which are found repeatedly in each others' company. They represent a tighter relationship than collocates, more like groups or phrases (but I call them clusters because these terms already have uses in grammar).

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*).

WordSmith Tools gives you two opportunities for identifying word clusters, in WordList and Concord. Both use the same method but **Concord** only processes concordance lines, while **WordList** processes whole texts.

How it does it…

Suppose your text begins like this:

Once upon a time, there was a beautiful princess. She snored. But the prince didn't.

If you've chosen 2-word clusters, the text will be split up as follows:

Once upon
upon a
a time
(note **not** "*time there*" because of the comma)
there was (etc.)

With a three-word cluster setting, it would send

Once upon a
upon a time
there was a
was a beautiful
a beautiful princess
But the prince
the prince didn't
(etc.)

That is, each n-word cluster will be stored, if it reaches n words in length, *up to a punctuation boundary*, marked by **;,.!?** (It seems reasonable to suppose that a cluster does not cross clause boundaries and these punctuation symbols help mark clause boundaries.)

Colours

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

plain text	this is the default colour
highlighted text	as above when selected
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 highlighted	concordance search word when selected
main sort word highlighted	first sort when selected

second sort word highlighted	first tie-breaker sort when selected
context word highlighted	context word when selected
most frequent collocate	most frequent collocate or detailed consistency word, p value in keywords
viewing texts	in the text viewer

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

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

See ▥ layout for changing the individual colours of each column of data.

Compute (π)

The point of it…

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

How to do it

Just press π 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.

Absolute and Relative

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

Examples

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

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

Abs(2;1) ÷ 5 for each row in your data, the new column will contain the data from column 2 of row 1, divide it by 5, and put the result in your new column. This example is just to illustrate; it would be silly as it would give the exact same result in every row.

Rel(2) ÷ Abs(2;1) × 100 for each row in your data, the new column will contain the data from column 2 of the same row, divide it by column 2 of row 1 and multiply it by 100, putting the result in your new column. This would give column 3 as a percentage of the top result in column 2. For the first row it'd give 100%, but as the frequencies declined so would their percentage of the most frequent item.

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

Copy

The quickest and easiest method of copying your data e.g. into your word processor is to select with the cursor arrows and then press Ctrl+Ins. This puts it into the clipboard.

If you press Copy you get various choices:

clipboard

saving as a text file

printing

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

all: copy all the rows and columns of data

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

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

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

Defaults: Permanent and Current

Settings can be seen by choosing *Review All Settings* (under *Settings*) in any Tool, and altered by choosing *Adjust Settings* in the WordSmith Tools Controller.

Any setting menu item in any Tool gives you access to these:

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

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

colours	customise the default colours
directories	set **WordSmith** so it "knows" which directories you usually use
text	character set, treatment of hyphens & numbers, default file extension
general	restore last file, printing, font
tags	tags to ignore, tag file
stop lists	for Concord, KeyWords and Wordlist
matching	files to match up, or lemma files to mark lemmas in a word list, etc.
Concord	number of entries, sort system, collocation horizons

KeyWords	procedure, max. p value, database & associate minimum frequencies
WordList	word length & frequencies, type/token #, cluster settings

permanent settings and wshell.ini file

You can save your settings by checking the save box after adjusting settings. Or by editing the **wshell.ini** file, installed when you installed **WordSmith Tools**. This specifies all the settings which you regularly use for all the suite of programs, such as your text and results directories, screen colours, fonts, the default columns to be shown in a concordance, etc.

Click on the file name to edit or view the following:

wshell.ini (in your \wsmith directory)
readme.txt (in your \wsmith directory)

sayings

Using **notepad,** you can edit sayings.txt (in your \wsmith directory), which holds sayings that appear in the main Controller window, if you don't like the sayings or want to add some more.

network and CD-ROM defaults

If you're running **WordSmith** straight from a CD-ROM, your defaults cannot be saved on it as it's read-only; Windows will find a suitable place for **wshell.ini**, usually the root directory of c:\.

The first time you use **WordSmith**, you will be prompted to Adjust Settings, choose appropriate Directories, Text Characteristics, Tag details etc. and enable the Save checkbox, after which your settings will be saved for future use. You can change settings and save them as often as you like.

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

restore last file (not available on network)

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 **Wshell** menu option or permanently in wshell.ini (in your \wsmith directory).

Directories

Found in main *Settings* menu in all Tools. Default directories can be altered in **WordSmith Tools** or set as defaults in **wshell.ini**.

Text Directory: where your text files are to be found.

WordList Directory: where you will usually save your word-list files.

KeyWords Directory: for your key-word list files.

Concordance Directory: for your concordance files.

If you write the name of a directory 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 directories and save files.)

tip

Use different directories 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 directory, but accessing them seems to slow down after there are more than about 500 in a directory. Use the batch facility to produce very large numbers of word list or key words files. I would recommend using a **\keywords** directory to store **.kdb** files, and **\keywords\genre1**, **\keywords\genre2**, etc. for the **.kws** files for each genre.

Editing Data (✓)

If this button is visible, then some of your data can be edited, e.g. the spelling of words in a word list. 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.

Pressing the button will bring up a window in which 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 the case of **WordList**, frequency changes will be updated in the Frequency window, though re-ordering will only take place if you re-order (F6), or after any unwanted entries have been zapped.

See also: Joining Entries

Editing Text File-names

As **WordSmith** makes concordances, word lists etc. based on a number of files, it keeps data on each text file as it goes along. If later, you want to view the source text files or do a concordance on them, the full drive and directory and file-name for each one can be retrieved from this.

If, however, your files have been moved, renamed, deleted, etc., **WordSmith** will not be able to find them. This is especially likely to happen in networks where the same text file may be **F**:\texts\one.txt for one workstation but **C**:\texts\one.txt for another.

For this reason, you may need to edit file-names, using the Replace button (⊕). Type in the search string (e.g. **f:**\) and the replacement (e.g. **c:**\), choose whether to confirm each change or not, and press OK.

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

Exiting

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 **Wshell** menu option or permanently in wshell.ini (in your \wsmith directory).

Filenames Button ☰

This button enables you to open a new window, displaying the text filenames from which your current data comes. You can edit these names if necessary (e.g. if the text files have been moved or renamed, or in a network, if they are on a drive whose name changes according to which machine you're running from.) To do so, press the Replace button (⊕).

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

Finding Relevant Files F

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 button, e.g. a key words listing.

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

What you get

A concordance based on all the words you marked, showing which text files they were found in. But it is a “fussy” concordance: any text files which doesn’t have *all* the words you selected get ignored.

Fonts

Found in main *Settings* menu in all Tools or via *Adjust Settings* | *General* 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. You can easily change the font size in the adjustment window here. To choose a different font name, press the Choose button, which opens up the standard Windows font choice dialogue box.

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.

By default any font chosen will be normal. To set a column of data to bold, italics, underline etc., use the layout button ▥.

WordSmith Tools will handle fonts to using other character sets (Cyrillic, Greek, etc.) if these have been installed, e.g. through Microsoft Plus in Windows 95.

You can set a permanent default font in your ***.ini*** file.

Layout & Format ▥

With a concordance, a word list or a key word list open, use this button to choose your preferred display formats for each column of data.

The top left window shows the available column headings. Click on one to activate it so that you can change these settings:

delete

press **Del** to delete a column of data. If you confirm the deletion, the whole column will be permanently deleted, e.g. after you've computed a new variable but no longer need it.

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.

typeface

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

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.)

width

in pixels. You can set this to the number of your choice. A standard monitor screen is 480 pixels wide, but better quality graphics cards and monitors allow much more than that.

decimals

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

colours

The bottom left window shows the available colours. Click on a colour to change the display for the currently selected column of information. The range of colour choices is determined by *Settings* in WordSmith Tools Controller.

Clipboard (Ctrl-C or Ctrl-Ins)

You can block an area of data, by using the cursor arrows and Shift, or the mouse, then press Ctrl-Ins 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.

The data will be copied to the Clipboard twice, using different formats:

a) as plain text, with a tab between each column. In this format no graphic data, such as Concord dispersion plots or KeyWords plots, can be included. The Windows plain text editor **Notepad** can handle only this data format. MicroSoft **Word** will paste (using Shift-Ins or Ctrl-V) the data as text. This format is needed if you want to re-use the data, e.g. in a spreadsheet.

In the case of a concordance line, the clipboard will copy the screen display, except that for the search-word to line up nicely, you will should use a non proportional font, such as Courier or Lucinda Console. If you don't like this, use the second option:

b) as a graphic, which includes screen colours and graphic data. MicroSoft **Word** can handle data in this format: see Edit | Paste Special | Picture. 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. Use this format for plotted data or in the case of a concordance, if you want only the words visible in your concordance line (not the whole line).

Notes

This allows you to jot down some notes to save with your data.

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

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

Printing and Print Preview (F3 and 🖶)

This takes you by default to a print preview, from which you can print.

Bigger and Smaller

Zoom in (🔍) or out (🔍), altering your view by 10% each time. 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.

Refresh

Refreshes (re-paints) the screen display.

Print Preview Menu

Black & White

Toggles monochrome output on and off. You can also set this in Printer Settings. Note that if black and white, colour is replaced by bold type, italics, or underlining.

Printer Setup

Standard Windows dialogue box, allowing you to change from portrait ▯ to landscape ▭, change printer or alter other printer settings. Also accessed in the Controller Printer Settings.

Print Current Page

This does what it says!

Other Pages etc.

Accesses the standard Windows dialogue box for printing. Here you can specify a range of pages to print, or choose to print to file.

Page Header

Gives you a chance to edit the header, which will be printed at the top of each page of the printout.

Print to File = Save As Text = Copy to a Text File

To save as text, choose the Print To File option, after pressing *Other Pages*. Or Copy 📋 to a text file.

See also : Printer Settings

Printer Settings

Found in *Adjust Settings* | *General* in the WordSmith Tools Controller.

Monochrome/Colour

If you set printing to black & white, 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.

Setup

Standard Windows dialogue box, allowing you to change from portrait ▯ to landscape ▭, change printer or alter other printer settings. You can also setup your printer in Print Preview.

Confirmation

You can set **WordSmith Tools** to confirm a print job in the defaults (wshell.ini) file. If this contains the line *confirm printing=YES* then every time you print you'll be told which lines of the current concordance or list were printed.,

See also : Printing

Saving (💾 and 💾)

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

Saved data is in a special **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 📋.

Save as Text (txt)

The point of it…

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

If you want to copy the data in colour, or export a plot, you should definitely use the clipboard.

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

How to do it

This function can be reached by txt or *Print to File* (via F3 or 🖨) or *Copy* (📋) to text file.

Options

These include:

header and footer	words you want to save at the start or end of the data (leave blank if none is wanted);
number each line	whether the numbers visible in the grey 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 to save	all/any which you have highlighted/a specific range, e.g. 1-10, 5-, -3 (row 0 is the grey one)
columns to save	all/any which you have highlighted/a specific range (column 0 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***.)

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.

Search for a Word, Suffix or Prefix (🔍 or F12)

All lists allow you to search for a word or part of one, or a number. The search operates on the *current column* of data. If you choose "Case sensitive" each letter must match in terms of upper case and lower case. You can choose to search "Down" from the next entry below the current one to the end, or else "Up" to the first entry.

Whole word – or bung in an asterisk

"Whole word" means search for a word with a word separator at each end. To search for a suffix or prefix, use the asterisk. Thus ***ed** will find any entry ending in **ed**; **un*** will find any entry starting with **un**. ***book*** will find any entry with **book** in it (**book**, **textbook**, **booked**, or **book** in a two or three word entry such as **she'll book it**.

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

See also: Searching by Typing, Search & Replace, Accented Characters & Symbols.

Searching 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: Searching for a word or part of one, Search & Replace, WordList sorting

Search and Replace ()

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

Whole word – or bung in an asterisk

"Whole word" means search for a word with a word separator at each end. To search for a suffix or prefix, use the asterisk. Thus ***ed** will find any entry ending in **ed**; **un*** will find any entry starting with **un**. ***book*** will find any entry with **book** in it (**book**, **textbook**, **booked**, or **book** in a two or three word entry such as **she'll book it**.

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

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

Match List (=)

The point of it…

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

How to do it

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

Text File or Template

Choose now whether you want to filter by using a text file which contains all the words you're interested in (e.g. a plain text file of function words [not supplied]) or a template filter such as ***ing** (which checks every entry to see whether it contains a word ending in **ing**.). If you choose a file, the Controller will then read it and inform you as to how many words there are in it.

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 with a tilde (~).

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

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

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 **edit.com** or **notepad**, which specifies all the words you wish to ignore. Separate each word using commas, or else place each one on a new line. Use capital letters. You can use square brackets to put comments in a stop list file. (Click here to see stoplist.stp (in your \wsmith directory) which you could use as a basis and save under a new name.)

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

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

See also: Making a Tag File

Suspending Processing

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

Continue

go on as if you had not interrupted anything

Finish this file, then stop

a graceful stop. Finishing the file means that you can keep track of what has been done and what there wasn't time for.

Stop now

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

Quit this Tool

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

Accented Characters & Symbols

This feature of Concordance Settings and other dialogue boxes enables you to insert symbols and accented characters into your search-word, exclusion word or context word, etc.

Just choose the symbol and drag it with your mouse to the place where you need it, or double-click on it.

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

See also: Choosing Language

Text Characteristics

Windows format etc.

You'll find a range of Windows and Dos and Internet character sets to choose from: you need the set used when these texts were first generated.

which files

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

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 as if they were ordinary words? If you set numbers off, words like *$300, 50.3M* or *10th* will be ignored in word lists, key words, concordances etc. (though they will still count towards word counts in text statistics).

Plain Text/HTML/SGML

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

start & end of heading

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

start & end of sentence

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

start & end of paragraph

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

characters within word

You may wish to allow certain characters within a word but not at either end of it. For example, 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.

See also: Tagged Text, Stop Lists.

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. With the exception of Viewer, and Concord, the main window for each Tool will appear to the right of it, and the same size. Each Tool main window will come just below any previous ones. Individual windows of results in each Tool will be restricted to that Tool's main window, and can be tiled or cascaded.

Make use of Alt-tab, which helps you to switch easily from one window to the next, and if using Windows 95, the *Start* toolbar.

minimising, moving and resizing windows

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

tile and cascade

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

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

screen clutter

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

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

restore last file

A convenience feature: the last file you saved or retrieved will by default be restored when you re-enter WordSmith Tools. I've kept it to one only to avoid screen clutter! This feature can be turned off temporarily via a **Wshell** menu option or permanently in wshell.ini (in your \wsmith directory).

Zapping ⚡

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

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

Tagged Text

Markup and Tags

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

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

You may wish to *select* within text files, e.g. cutting out a header or getting only the conclusions, instead of using the whole text. And you might want to get **WordSmith Tools** to choose only files meeting certain criteria, e.g. having "sex=f" in a text file header section, where the speaker is a woman.

See also: Handling Tags, Making a Tag File, Showing Nearest Tags in Concord, Types of Tag, Viewing the Tags, Using Tags as Text Selectors

Types of Tag Markup

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.

six special markers

Six kinds of marker may be marked as significant for word lists: those which represent starts and ends of headings, 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 can use the same opening and closing symbols, usually some kind of brackets, for all your tags (as the British National Corpus does using SGML markup): **<Noun>,<Verb>,<Pronoun>**, or you can use a mixture: **<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;t&eacute*** which represents **été.**

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

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

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

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

See also: Overview of Tags, Handling Tags, Making a Tag File, Showing Nearest Tags in Concord, Viewing the Tags, Using Tags as Text Selectors

Handling Tags

ignore all tags

Specify all the opening and closing symbols in *Adjust Settings |Tags |Tags to Ignore* and such tags will be simply left out of word lists and concordances, as if they weren't in the original text files. example : **<*>**   this will cut out all wording starting at a < symbol and ending at a > symbol (up to 1,000 characters).

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 (up to 1,000). 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 the checkbox beside the filename is checked. (This is to allow you to swap options easily without having to retype filenames.) If active, when you press OK, you will see a message saying how many tags have been read in from the tag file. You can see the individual tags under *Review All Settings*.

translate entity references into other characters

If you use SGML or HTML tagged text, you may want to translate symbols. For example, SGML uses **&mdash;** instead of a long dash. To do this, first prepare a Tag File which contains the strings you want to translate. Then choose *Adjust Settings | Tags | Tag File 2 (tags to be translated)* and choose your tag file. **WordSmith** will then translate any entity references in this file into the corresponding characters.

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

See also: Overview of Tags, Making a Tag File, Showing Nearest Tags in Concord, Types of Tag, Viewing the Tags, Using Tags as Text Selectors

Using Tags as Text Selectors

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

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

selecting within texts (only part of file)

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; the default is 0.

Sections to Cut

If you are using text files with SGML, XML or HTML headers (e.g. the British National Corpus) you may well want to cut out the header from your word lists, concordances, etc.

To do so, specify what is to be cut, starting at (by default **start of file**, but you could type for example **<HEAD>**) and where you want to cut to (e.g. **</HEAD>**). You can choose to cut out up to 2 different and separate sections (**<HEAD>**,**</HEAD>**;**<BODY>**,**</BODY>**). Unless you specify **start of file**, this function cuts out any section located as many times as it is found within the whole text.

This function assumes each section to cut is less than 30,000 characters long. You will need to check the activate box below it, too.

Sections to Use

Here you want to select one section of a text and ignore the rest. To do so, specify one tag to define the desired start, and one to specify the end, e.g. **<Intro>** and **<Body>** (these would analyse only text introductions), or **Mary:** and **Peter:** (these would get all of Mary's contributions in the discourse but nothing else).

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

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

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

This function depends on your text containing the required tags of course; also, it will only work if the gap between the start and the end is smaller than 30,000 characters.

selecting between texts (only if containing)

This function allows you to search through the first 30,000 characters of each text, e.g. in text headers. 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 the first 30,000 characters of every text file (not just the tags or header sections) to check that it meets your requirements.

You can specify up to 12 tags, each up to 30 letters 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.

You will get a chance to choose to save the set of files which meet your requirements for later use as favourites.

Examples:

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

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

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 Prize* in the first two boxes in the bottom row.

the order in which these choices are checked

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. Delete any section specified as a "section to cut";

3. If there are "sections to use", delete everything which is not within them;

4. Cut start of each line, if applicable;

5. Process any entity references you want to translate;

6. Delete any tags to ignore (those not to be retained).

defaults

The defaults are: select *all* sections of *all* texts selected in Choose Texts.

See also: Overview of Tags, Making a Tag File, Tag Handling, Showing Nearest Tags in Concord, Viewing the Tags, Types of Tag

Making a Tag File

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

Up to 1,000 pre-defined tags can be stored. They will be case sensitive.

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

A tag file for tags to retain contains a simple list of all the tags you want to retain. Sample tag list files for MicroConcord Corpora and BNC handling (click here to see mconc.tag or bnc.tag and bnclex.tag) are included with your installation (in your \wsmith directory): you could make a new tag file by reading one of them in, altering it, and saving it under a new name.

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

> **&Eacute; É**
>
> **&eacute; é**

A sample tag file for translation (click here to see sgmltrns.tag) is included with your installation (in your \wsmith directory): 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, Types of Tag, Viewing the Tags, Using Tags as Text Selectors

Concord Index

Explanations

Settings

Procedures

What is *Concord* and what's it for?

Concord is a program which makes a concordance using DOS, Text Only, ASCII or ANSI text files.

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

The point of it…

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

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

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

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

Concordance - What you can see and do

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

The Columns

These show the details for each entry: the entry number, the concoirdance line, set, tag, word-position (e.g. 1st word in the text is 1), source text filename, and how far into the file it comes (as a percentage).

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 clear, press 0 (zero).

Tag

This column shows the nearest tag.

Stretching the display to see more

You can pull the concordance display to widen its column. Just place the mouse cursor on the grey bar between one column and another; when the cursor changes shape you can pull the whole column. The same applies to each individual row: place the mouse cursor between one row and another in the grey numbered area, and drag.

Or press ■ to "grow" all the rows, or ▬ to shrink them.

Viewing the original file (🗒)

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

See also:
Re-sorting your concordance lines
User-defined categories
Altering the View
Blanking out the search-word
Collocation (words in the neighbourhood of the search-word)
Plot (plots where the search-word came in the texts)
Clusters (groups of words in your concordance)
Editing the concordance
Zapping entries
Saving and printing
Window Management

Altering the View of a Concordance

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

Sentence Only

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

Tags and Spaces Cut

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

See also: showing nearest tags.

Blanking out the search-word

Blanking Out Concordance Entries

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.

Collocation

What's a "collocate"?

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

The point of it...

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

Options

You may compute a concordance with or without collocates: without is slightly quicker and will take up less room on your hard disk. The default is to compute *with* collocates. The number of collocates stored will depend on the collocation horizons.

Collocates can be viewed after the concordance has been computed.

Technical Note

The literature on collocation has never distinguished very satisfactorily between collocates which we think of as "associated" with a word (*letter* - *stamp*) on the one hand, and on the other, the words which do actually co-occur with the word (*letter* - *my, this, a*, etc.).

We could call the first type "coherence collocates" and the second "neighbourhood collocates" or "horizon collocates". It has been suggested that to detect coherence collocates is very tricky, as once we start looking beyond a horizon of about 4 or 5 words on either side, we get so many words that there is more noise than signal in the system.

KeyWords allows you to study Associates, which are a pointer to "coherence collocates". **Concord** will supply "neighbourhood collocates". **WordList** allows you also to study Mutual Information

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.

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

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

The plot is initially sorted by no. of words per 1,000.

There are two ways of viewing the plot, as a *Uniform Plot* (the default, checked, where all plotting rectangles are the same length) or not (unchecked, where the plot rectangles reflect the original file size -- the biggest file is longest). Change this in the *Settings* menu at the top.

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.

Each plot window is dependent on the Concordance from which it was derived. If you close the original concordance down, it will disappear. You can *Print* the plot. There's no *Save* option because the data come from a Concordance which you should Save, or *Print to File*. You can *Copy* to the clipboard (Ctrl-Ins) and then put it into a word processor as a graphic, using Paste Special.

Re-sorting Dispersion Plot (⟳ or F6)

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

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>
```

You can sort on tags.

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.

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

Clusters in Concord (/../)

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. Another feature in **Concord** which helps you see patterns is Patterns.

How it does it…

When you press /../, **Concord** goes through your current concordance lines, seeking the repeated word clusters.

Clusters are sought within the limits you set in the menu under *Settings* | *Clusters*: default: 5 words left and right of the search word, but up to 25 left and 25 right allowed. These settings allow you to choose how many words a cluster should have (cluster size 2 to 4 words recommended), and how many of each must be found for the results to be worth displaying (say 3 as a minimum frequency).

Clusters will stop at clause boundaries as signalled by these six punctuation marks **;,!?:.** Naturally, they will often contain the search-word itself, since they are based on concordance lines.

It's a dependent window

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

See also: general information on clusters, WordList Clusters.

Concordance Settings

Search Word or Phrase

Type the word or phrase **Concord** will search for when making the concordance, or the name of a file of search words. You may also choose from a history list of your previous search words. For details of syntax, see Search Word Syntax.

Excluded Word(s)

You can also specify one or more "exclusion words", using the same syntax. If the search word is *book** and the exclusion word is *booked*, you'll get *book*, *books*, *booking*, *bookable*, but not *booked*. This is only useful (and only visible) if you've used a wild-card (**?** or *****) in the search word.

Context Word(s) & Context Search Horizons

Symbols

A set of symbols and accented characters to use, based on current Text Characteristics, which you can insert into your search word, context word, or exclusion word. To insert, either drag the one you need to the search word, context word, or exclusion word, or else double-click on it. The symbol you've chosen will be inserted into the word at the current insertion point in the usual Windows style (replacing any highlighted section).

Other settings affecting a concordance are available too; see WordSmith Controller Concordance Settings;Accented characters; Choosing Language

Search Word Syntax

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

Examples

Search Word	Finds
book	*Book* or *book* or *BoOk*
book	*book, books, booking, booked*
***book**	*textbook* (but not *textbooks*)
bo* in	*book in, books in, booking in* (but not *book into*)

book * hotel	*book a hotel, book the hotel, book my hotel*
bo* in*	*book in, books in, booking in, book into*
book?	*book, books, book; book.*
book^	*book, books*
b^^k	*book, back, bank,* etc.
==book==	*book* (but not *BOOK* or *Book*)
book/paperback	*book* or *paperback*

symbol	meaning	example
*	disregard the end of the word	*book**
?	any single character (including punctuation) will match here	*Engl???*
^	any single letter of the alphabet will match here	*Fr^nc^*
==	case sensitive	*==French==* *==Fr*==*
:\	means use a file for up to 500 search-words (see file-based search_words)	*c:\text\frd.txt*
/	separates alternative search-words. You can specify up to 15 alternatives within an 80-character overall limit	*may/can/will*

If you want to use ***, ? , == , ^ , :** or / as a character in your search word, put it in double quotes. Examples:

"*"
Why"?"
and"/"or
":\"

see also: Context Word

WordSmith Controller Concordance Settings

These are found in the main Controller under *Adjust Settings* | *Concord.*

This is because some of the choices -- e.g. collocation horizons -- may affect other Tools.

Entries Wanted

The maximum is 16,368 lines. This feature is useful if you're doing a number of searches and want, say, 100 examples of each. In that case, the 100 entries will be the first 100 found.

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

Characters in 'save as text'

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

Sort

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

Collocates

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

See also Concord Help Contents.

Editing a Concordance

The point of it…

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

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

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

See also: Window Management

Saving and Printing a Concordance

You can save the concordance (and its collocates 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.

File-based Search Words or Phrases

The point of it…

As **Concord** allows up to 16,368 entries altogether, 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…

Use a Windows editor (e.g. **Notepad**) to prepare this file. A sample (**concsws.txt**) is included with the distribution files. The limit is 500 words or phrases. Each one must be on a separate line of your file.

Then, instead of typing in each word or phrase in the Search Word dialogue box, just type in the file name, e.g. **c:\wsmith\concsws.txt**.

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 **\wsmith** directory unless you're running on a network in which case it'll be in Windows' temporary directory, e.g. **\windows\temp**. It's up to you to delete the temporary file.

(By default, **Concord** will assume that a search-word including **:** should use this file-based procedure and will do so if it finds the file. So if you want to include **:** in your search-word and don't want it to refer to a file, put **":\"** instead.)

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 the context word. This will only find *book* if *paper* or *papers* is **not** nearby.

The context horizons determine how far Concord must look to left and right of the search word when checking whether the search criteria have been met. The default is 5,5 (5 to left and 5 to right of the search word) but this can be set to up to 25 on either side. 0,2 would look only to the right within two words of the search word.

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

Syntax is like that of the search word or phrase,

* means disregard the end of the word and can be placed at either end of your context word.

== means case sensitive

~ means not this word, e.g. *~swim* means *swim* may **not** be in the context specified.

/ separates alternatives. You can specify up to 15 alternatives within an 80-character overall limit. If you have various alternatives, any negative ones (with ~) are checked first.

If you want to use ***, ? , == , ~ , :**\ or / as a character in your search word, put it in double quotes, e.g. **"*"**

Collocate Horizons

The collocate horizons represent the number of collocates **Concord** will find to the left and right of your search word, and the distance used by **KeyWords** in searching out plot-links. The default is 5,5 (5 to left and 5 to right) but you can go up to 25 on either side.

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

see also: Collocate Settings

Collocation Settings

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

Minimum Specifications

The minimum length is 1, and minimum frequency is 1. 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 force storage to occur only if the collocate comes 5 or more times in the neighbouring context.

Similarly, you can specify how long a collocate must be for it to be stored in memory, e.g. 3 letters or more would be 3.

Horizons

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

Collocation Display

The collocation display initially shows the collocates in frequency order. Beside each word, you'll see 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:

1. the total number of times the word was found in the neighbourhood of the search word
2. the total number of times it came to the left of the search-word
3. the total number of times it came to the right of the search-word
4. a set of individual frequencies to the left of the search word (5L, i.e. 5 words to the left, 4L .. 1L)
5. an asterisk column, representing the search-word
6. a set of individual frequencies to the right of the search word (1R, 2R, etc.)

The number of columns in (2) and (4) will depend on the collocation word horizons. With 5,5 you'll get five columns to the left and 5 to the right of the search word. So you can see exactly how many times each word was found in the general neighbourhood of the search word and how many times it was found exactly 1 word to the left or 4 words to the right, for example. The most frequent will be signalled in red. The frequency display can be re-sorted () and you can recalculate the collocates () if you zap entries from the concordance or change the horizons.

The point of all this is to work out characteristic lexical patterns. It can be hard to see overall trends in your concordance lines, especially if there are lots of them. By examining collocations in this way you can see common lexical and grammatical patterns of co-occurrence. Collocational linkages which involve grammatical items are often referred to as *colligation*.

see also: Collocation, Mutual Information

Re-sorting Collocates (or F6)

The frequency-ordered collocation display can be re-sorted to reveal the frequencies sorted by their *total* frequencies overall (the default), by the left or right frequency total, or by any individual frequency position, from 25 words to the left of the search word to 25 words to the right. The sort can be either ascending or descending, the default being descending.

The point of it…

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

The choices depend on the collocation horizons.

See also: Collocation, Collocation Display

Re-sorting a concordance (or F6)

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

The point of it…

The point of re-sorting is to find characteristic lexical patterns. It can be hard to see overall trends in your concordance lines, especially if there are lots of them. By sorting them you can separate out multiple search words and examine the immediate context to left and right. For example you may find that most of the entries have "in the" or "in a" or "in my" just before the search word -- sorting by the second word to the left of the search word will make this much clearer.

Sorting is by a given number of words to the left or right (L1 [=1 word to the left of the search word], L2, L3, L4, L5, R1 [=1 to the right], R2, R3, R4, R5), on the search word itself, the context word (if one was specified), the nearest tag, the distance to the nearest tag, a set category of your own choice, or original file order (file).

Main Sort

The listing can be sorted by three criteria at once. A *Main Sort* on Left 1 will sort the entries according to the alphabetical order of the word immediately to the left of the search word. A second sort (… *then by* ...) on Right 2 would re-order the listing by tie-breaking, that is: only where the Left 1 words (immediately to the left of the search word) matched exactly, and would place these in alphabetical order of the words 2 to the right of the search word. For very large concordances you may find the third sort (*...finally by*) useful: this is an extra tie-breaker in cases where the second sort matches.

For many purposes tie-breaking is unnecessary, and will be ignored if the first and second sorts are the same (e.g. Left 1 and Left 1) or if the "activated" box is not checked.

sorting by set (user-defined categories)

You can also sort by set, if you have chosen to classify the concordance lines according to your own scheme, using letters from **A** to **Z** or **a** to **z**. The sort will put the classified lines first, in category order, followed by any unclassified lines (which will appear in a light grey colour). See Nearest Tag for details of sorting by tags.

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

Case Sensitivity

Case sensitivity is set in the re-sort dialogue box. If it's off you will get

brother
Brother
brother

If it is checked you'll get capitalised words at the top:

Brother
brother
brother

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**.

À with A, Ç with C

Accented characters are by default treated as equivalent to their unaccented counterparts in some languages (so, in French we get **donné, donner, donnes, 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** before **cas**)

If this box is checked, sorting will treat accented characters as related to their unaccented counterparts according to the current language.

Punctuation

Concord takes punctuation into account when sorting:

brother
brother!
brother,
brother.
brother;
brother's
brothers
brothers'

If the apostrophe is one of the Characters within Word the item **brother's** will come just before **brothers** as shown above; otherwise it will follow **brother!** -- the reason has to do with character code sequences. If you're working in English you'll probably want to keep the apostrophe in the Characters within Word.

See also: Accented characters; Choosing Language

User-defined 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. You can have up to 52 categories. To allocate a concordance line to a letter between **a** and **z**, just press the key. With Shift, you get **A** to **Z**.

You can later sort the concordance lines using these categories; the upper-case ones will all come before any in lower-case.

You can also save only the entries which are in one of these categories; to do this, choose Copy, then **.CNC** and **specify** (or **selected** to save only the highlighted entries).

To clear, press **0** (zero).

Patterns (↑)

When you have a collocation window open, one of the button and menu options is *Patterns*. This will show the words adjacent to the search word, organised in terms of frequency within each column. That is, the top word in each column is the word most frequently found in that position. The second word is the second most frequent.

The 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.

KeyWords Index

Explanations

Settings and Procedures

Tips

Definitions

What is *KeyWords* and what's it for?

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.

Key Words Example

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.

WordSmith Controller KeyWords Settings

These are found in the main Controller under *Adjust Settings | KeyWords*.

This is because some of the choices may affect other Tools. **KeyWords** and **WordList** both use similar routines: **KeyWords** to calculate the key words of a text file, and **WordList** when comparing comparing word-lists.

Procedure

The default is Log Likelihood. See procedure for further details.

Max. p value

See p value for more details.

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

You may want to restrict the number of key words 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 hardly likely to be "key". The default setting of 3 mentions as a minimum helps reduce spurious hits here, though in the case of short texts, e.g. less than 500 words long, you may wish to change the minimum to 2.

Database: minimum frequency

The default is 1. See database.

Database: associate minimum frequency

The default is 5. See associates.

See also KeyWords Help Contents.

Definition of Key-ness

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

A word is said to be "key" if

a) it occurs in the text at least as many times as the user has specified as a Minimum Frequency

b) its frequency in the text when compared with its frequency in a reference corpus is such that the statistical probability as computed by an appropriate procedure is smaller than or equal to a p value specified by the user.

positive and negative keyness

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

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

typical key words

KeyWords will usually throw up 3 kinds of words as "key".

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

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

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

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

How Key Words are Calculated

The "key words" are calculated by comparing the frequency of each word in the smaller of the two wordlists (the one shown on the left) with the frequency of the same word in the reference wordlist. All words which appear in the smaller list are considered, unless they are in a stop list.

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

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

its frequency in the small wordlist

the number of running words in the small wordlist

its frequency in the reference corpus

the number of running words in the reference corpus

and cross-tabulates these.

Statistical tests include:

the classic chi-square test of significance with Yates correction for a 2 X 2 table

Ted Dunning's Log Likelihood test, which gives a better estimate of keyness, especially when contrasting long texts or a whole genre against your reference corpus.

A word will get into the listing here if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger wordlist.

Unusually *in*frequent 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.

2 Word-list Analysis

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

Choose Word Lists

In the dialogue box you will choose 2 files. Change directories at each side if necessary, then highlight the text file in the left box and the reference corpus file in the right box.

See also How Key Words are Calculated, KeyWords Settings

Batch Processing of Key-word Lists (Control-B)

KeyWords can also carry out a series of analyses, using one reference corpus word-list file and comparing it with a whole series of word lists based on smaller text files. These word lists may have been made as a batch or separately.

Choose the small word lists in the list on the left, by clicking on them, pressing Control (to add one more) or Shift (to add a range). In the right dialogue box you will choose 1 reference corpus file.

The point of it…

It's faster in a batch. Batch processing is the same as making a databases, if you leave the option to store as a database checked; in this case the batch will take up less disk room too.

See also: WordList Batch Processing, Batch File-names, KeyWords Settings

p value

The ***p*** value is that used in standard chi-square and other statistical tests. This value ranges from 0 to 1. A value of .01 suggests a 1% danger of being wrong in claiming a relationship, .05 would give a 5% danger of error. In the social sciences a 5% risk is usually considered acceptable.

In the case of key words analyses, where the notion of risk is less important than that of selectivity, you may often wish to set a comparatively low ***p*** value threshold such as 0.000001 (one in 1 million) so as to obtain fewer key words. Or you can set a low "maximum wanted" number in the main Controller, under *Adjust Settings | KeyWords*.

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

See also: Definitions

Key Words display

The display shows

1. each key word

2. its frequency in the source text(s) which these key words are key in, *italicised*.

3. the name of the source text file (or the word list file name if there's more than one) and **%**, also in *italics*.

4. its frequency in the reference corpus

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

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

7. p value.

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

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

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

view button

This enables you to see the original source text using Viewer, and will highlight the key words.

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

Keywords Plot ()

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.

see also: Plot Links, Key words plot display

Key Words Plot Display

The plot will give you useful visual insights into how often and where the different key words crop up in the text. The plot is initially sorted to show which crop up more at the beginning (e.g. in the introduction) and then those from further in the text.

view button

This enables you to see the original source text using Viewer, and will highlight the key words.

re-sorting

You can re-sort the listing using . Re-sorting rotates through the following types:

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

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.

in

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, double-click on the word you're interested in. A new window will open up, 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. Links windows should help you gain insights into the lexical relations here.

Each plot window is dependent on the key words listing from which it was derived. If you close that down, it will disappear. You can *Print* it. There's no *Save* option because the plot comes from a key words listing which you should *Save*, or *Save As*. There's no save as text option because the plot has graphics, which cannot adequately be represented as text symbols, but you can *Copy* to the clipboard (Ctrl-Ins) and then paste it into a word processor as a graphic. Alternatively, use the *Output* | *Data as Text File* option, which saves your plot data (each word is followed by the total number of words in the file, then the word number position of each occurrence).

The ruler button (📏) 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

Plots and Key-word Links

The point of it…

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

A plot will show where each key word occurs in the original file. It also shows how many links there are between key-words.

What are links?

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

Suppose the word *elephant* is key in a text about Africa, and that *water* is also a key word in the same text. If *elephant* and *water* occur within a span of 5 words of each other, they are said to be "linked". The number of times they are linked like this in the text will be shown in the Links window.

What you see

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

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

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

Re-sorting in KeyWords (⟳or F6)

A **key words list** offers a choice between sorting by

key-ness	(the *key*est words appear at the top)
alphabetical order	(from A to Z)
frequency in the smaller list	(the most frequent words come first)

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

A **key words plot** rotates between sorting by

key-ness	(the *key*est words appear at the top)
alphabetical order	(from A to Z)
frequency	(words which appear oftenest come first)
number of links	(the most linked words come first)
first mention of each key word in the text	
range	(words used in smallest sections of text come first)

A **key key words database** toggles between sorting by

frequency	(the most *key key* words appear at the top)
alphabetical order	(from A to Z)

An Associates 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)

Calling up a Concordance

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

The point of it…

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

Converting your data into a word list

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).

Database of Key Key-words

(default file extension .KDB)

The point of it…

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

For example, if you have 500 business reports, each one will have its own key words. These will probably be of two main kinds. There will be key-words which are key in one text but are not generally key (names of the firms and words relating to what they individually produce); and other, more general words (like *consultant*, *profit*, *employee*) which are typical of business documentation generally.

By making up a database, you can sort these out. The ones at the top of the list, when you view them, will be those which are most typical of the genre. The list is ordered in terms of "key key-ness" but can be toggled into alphabetical order and back again.

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

Definition of a Key Key-word

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

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

Creating a Database (Control-D)

To build a key words database, you will need

a) a set of texts in plain text format. The database is likely to be most useful if there are a lot of texts and if they are selected in some principled way in terms of genre. For example, you might have 250 student compositions, or 300 business reports.

b) a set of corresponding word-lists. These should be made using **Wordlist**: for further details see the **Wordlist** help file. The default file extension is .LST. The easiest way to make them, however, will be using Wordlist's batch file mode (default file extension .WDB). The word-list files should all be in the same directory.

c) a reference word-list file, also made using **Wordlist**. This could be a single word-list (extension .LST) consisting of all the words in your set of texts. In this case you would have 250 individual word-lists of all your student compositions, plus another big one consisting of all the words in all 250 compositions.

Alternatively, you might make up a reference word-list file based on some other texts against which you want to compare the compositions. For example, a word-list based on a number of literary works or essays by native speakers. (If all your student compositions were on the topic "A day at the market", the word *market* would not appear as a key-word if the compositions were individually compared against the set of all the compositions; if they were compared with a general native-speaker set of essays, the word *market* would come up as key. In other words, what is "key" is key by comparing frequencies with some norm or other and you should decide what norm you want to use.)

In **KeyWords**, in the main menu, select *New Database*. This enables you to choose the whole set of word-list files, and to specify which reference file to compare them against, as **KeyWords** proceeds, file by file. It also enables you to choose the minimum frequency and degree of outstandingness required (the p value). Note that making a database means that only positive key words will be retained.

Once you've chosen your files, you'll see a dialogue box headed *Batch of Files*.

First specify where the resulting database is to be put. You can specify a new sub-directory and if necessary it'll be created.

Second, specify the filename(s). If you are making a single database (the default), only one file will be created. Otherwise you can choose for the filenames to be either a) brand new ones based on a mask of your choice, or b) a unique combination of the filename of each word-list, and the first two letters of the name of the reference word-list file.

After you've made the database, you can then Open it to see the key key-words and associates.

Associates

"Associates" is the name given to key-words associated with a key key-word.

The point of it…

The idea is to identify words which are commonly associated with a key key-word, because they are key words in the same texts as the key key-word is. An example will help.

Suppose the word *wine* is a key key-word in a set of texts, such as the weekend sections of newspaper articles. Some of these articles discuss different wines and their flavours, others concern cooking and refer to using wine in stews or sauces, others discuss the prices of wine in a context of agriculture and diseases affecting vineyards. In this case, the associates of *wine* would be items like *Chardonnay, Chile, sauce, fruit, infected, soil*, etc.

The listing shows associates in order of frequency. A menu option allows you to re-sort them.

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

See also: definition of associate.

Definition of Associate

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

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

Regrouping the 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

Choosing Word List Files

You'll see 2 lists of the existing word list files in the default results directory, headed *Word List(s)* and *Reference Corpus List*.

If you have chosen *Batch* analysis, you'll be able to choose as many word list files as you like in the left display (press Control as you click to select non-adjacent lists, or Shift to select a range). Then choose **one** file for your reference corpus from the other side.

If you have chosen to *make a single keyword list*, you must choose **one** file from **each** list.

No word-lists visible

If you can't see any word lists in the displays, either change directories 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.

KeyWords Advice and Tips

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!

Word-List Index

Explanations

Procedures

Zapping entries

see also : WordSmith Main Index

What is *WordList* and what's it for?

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

The point of it…

These can be used

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

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

These word-lists may also be used as input to the KeyWords program, which analyses the words in a given text and compares frequencies with a reference corpus, in order to generate lists of "key-words" and "key-key-words".

WordSmith Controller WordList Settings

These are found in the main Controller under *Adjust Settings* | *Wordlist*.

This is because some of the choices -- e.g. Minimum & Maximum Settings -- may affect other Tools.

Word Length & Frequencies

See Minimum & Maximum Settings.

Type/Token

See WordList Type/Token Information.

Clusters

See Single Words & Clusters.

Case Sensitivity

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

See also Using Index Lists, Viewing Index Lists, WordList Help Contents.

Using Index Lists

the point of it

One of the uses for an Index List is to record the positions of all the words in your text file, so that you can subsequently see which word came in which part of each text. Another is to speed up access to these words, for example in concordancing. If you select one or more words in the index and press C, you get a speedy concordance.

Another is to compute "Mutual Information" scores which relate word types to each other.

Thirdly, you may wish to record an index of the different words as they appear in your text(s). Such a list will therefore not be in frequency or alphabetical order, but "file order".

See also Making an Index Lists, Viewing Index Lists, WordList Help Contents.

Making an Index List

To create an index, first use the main Controller and choose *Adjust Settings* | *Indexing*.

You will need to

- specify: the file name for the index;
- activate it;
- (optionally) activate File Order too;
- choose a High Frequency cut-off number (explained below).

stages

The index is created in 3 stages. In the first stage, **WordList** will go through your selected texts and identify all the word types; in the second pass, it processes only the high frequency items (those types which can be expected to occur at least say 500 times. 500 is a reasonable "high frequency number"). Then in the third pass it will process the remaining word types found. So to enable stage 2, you will need to choose a suitable number (which may be affected by the amount of memory your machine has).

These stages are needed because **WordList** is really working hard. It is having to store information about the position of every instance of every word-type! In the case of a corpus of many millions of words, that stretches the limits even of a modern pc. (I am using a Pentium 200 with 64MB of RAM which is not by any means a super-equipped machine nowadays. It makes a full index of the 2 million tokens in the BNC Sampler in about 8 minutes. 10 million words of BNC spoken English, reading from a 20 speed CD-ROM, takes 30 minutes.)

index file types and sizes

Two or three files may be created for each index:

.wdx file: a small file containing only settings and text file data.

.xfo file: (if *File Order too* is activated) knows the file order position of each word token. This file is about half the size of your original corpus and permits the computation of Mutual Information scores for each word type.

.xal file: knows the position of all the instances of each word (in alphabetical order); can show them to you and enable very quick concordancing. This **.xal** file will be slightly larger than the size of your original corpus.

See also Using Index Lists, Viewing Index Lists, WordList Help Contents.

Viewing Index Lists

In **WordList**, the menu options *Index | Alphabetical List* and *Index |F ile Order List* give you two kinds of access to the index.

alphabetical

This shows the words in alphabetical order. Each word type is followed by the number of occurrences of each type (Freq.). Next you will see a small piece of concordance line for each word type, actually showing the first mention of the word in a minimal context. If mutual information has been computed, the remaining columns show each related word, and its Mutual Information score. The headings go from Related 1 up to Related 10 for a maximum of 10 associated word types.

You can highlight one or more words or mark them with the *m* button, then press **C** to get a speedy concordance.

file order

The file order list shows each word-type as it appeared in the text(s), with a plot of the occurrences.

Depending on the speed of your pc, the plot may seem slow in displaying in the case of the most frequent items, which will usually start to appear fairly near the top of the list. You can speed things up if necessary by choosing to make the plot invisible using the ▥ layout button.

You could save this list as text, thus making an index of your text(s).

See also Making an Index Lists, WordList Help Contents.

Mutual Information (◈)

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 besides *problem* than solve, so the procedure for calculating Mutual Information takes into account not just the most frequent words found near the word in question, but also whether each word is often found elsewhere, well away from the word in question. Since *the* is found very often indeed far away from *problem*, it will not tend to be related, that is, it will get a low MI score.

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

See Oakes for further information about Mutual Information.

Settings

The Mutual Information settings are found in the Controller under *Adjust Settings* | *Indexing* or in a menu option in **WordList**.

procedure: for the moment, only MI. This uses the formula (joint frequency × corpus size in tokens) ÷ (frequency of node × frequency of collocate).

Log base 2: computes MI score according to the formula, then takes the logarithm to the base 2 of it.

ignore sentence, para & heading breaks: if checked, does not stop considering collocates up to the horizons specified; if not checked (the default), it will stop at the horizon limits or else at a sentence, heading or paragraph boundary, whichever is the earlier.

max. percent ignores any tokens which are more frequent than the percentage indicated. (The point of this is to avoid computing mutual information for words like the and of, which are likely to have a frequency greater than say 1.0%.)

min. mutual info. the minimum number which the MI must come up with to be reported. If Log2 is enabled, a useful limit is 3.0. Below this, the linkage between node and collocate is likely to be rather tenuous.

min. frequency: the minimum frequency for any item to be considered for the mutual information calculation (default = 4). (If an item occurs only once or twice, the mutual information is unlikely to be informative.)

The word list index shows up to 10 collocates of each node, arranged in order of mutual information, together with the MI score for each.

See also Computing Mutual Information, Making an Index Lists, Viewing Index Lists, WordList Help Contents.

Computing Mutual Information ( )

To compute Mutual Information (MI) you need a WordList Index, made with "File Order too" activated. Call up the alphabetical view of the list.

If no entries are selected (highlighted) when you press , MI will be computed for each item as long as its basic frequency is at least as high as the minimum frequency required in MI Settings.

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

Computing the MI score for each and every entry in an index takes a long time: at least a couple of hours in the case of an index based on 10 million words.

Don’t forget to save your results afterwards.

See also Mutual Information Settings, Making an Index Lists, Viewing Index Lists, WordList Help Contents.

Batch Processing of Word Lists

one by one v. batch mode

A set of 10 texts could be made into one big word list, giving the frequencies of each word as a percentage of the whole set of 10 texts, or into 10 separate word lists, giving the frequencies of each word as a percentage of the individual text.

the point of batch processing

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

If you have selected more than one text file you can ask **WordList** to process them all as a batch: one word list per text file. This is an advantage when preparing for **KeyWords** databases.

The batch of word lists can be stored as separate word list files, or stored all within one big computer file (a database). The advantage of this database strategy is that it wastes less hard disk space; the disadvantage is that it's trickier to access the individual word lists within the big file. If you're preparing word lists for a **KeyWords** database, it may be best to store it in a database.

You are asked to choose a suitable name as a basis for the batch of files. This can be either based on the original file names or else on a mask.

See also: KeyWords Batch Processing

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: 14 letters. You can allow for words of up to 50 characters in length. If a word exceeds the limit and ***Abbreviate with*** + is checked, **WordList** will append a + symbol at the end of it to show that it was cut short. (If ***Abbreviate with*** + is not checked, the long word will be omitted from your word list. You might wish to use this to set both minimum and maximum to say, 4, and leave ***Abbreviate with*** + un-checked – that way you'll get a wordlist with only the 4-letter words in it.

minimum frequency

Default: 1. By default, all words will be stored, even those which occur once only. If you want only the more frequent words, set this to any number up to 32,000.

maximum frequency

Default maximum is 2,147,483,647 (2 Gigabytes). You'd have to analyse a lot of text to get a word which occurred as frequently as that!. You might set this to say 500, and the minimum to 50: that way your word-list would hold only the moderately common words.

type/token mean number (default 1,000)

Enables a smoothed calculation of type/token ratio for word lists. Choose a number between 10 and 20,000. For a more complete explanation, see WordList Type/Token Information.

See also: Text Characteristics, Stop Lists, Setting Defaults

Case-sensitive Word Lists

Normally, you'll make a case-insensitive word list, especially as in most languages capital letters are used not only to distinguish proper nouns but also to signal beginnings of sentences, headings, etc. If, however, you wish to make a word list which distinguishes between *major*, *Major* and *MAJOR*, activate case sensitivity (Adjust Settings | WordList | Case Sensitivity in the Controller).

You will also need to define and choose a language which includes lower-case letters. Use one of the "Other" languages for this.

Warning: if you choose to define your own language features, keep a backup safe. The alphabet defined in it will be needed to view any lists you save when it is in use.

Single words v. Clusters in WordList

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.

How to do it…

To activate **WordList** cluster processing, choose the option *Settings* | *Min. & Max. Frequencies* in the main WordList menu. Choose a cluster size from 1 (i.e. single words) to 8.

After frequency sorting you will see the most characteristic clusters, where the frequency is at least 2 (since those which come up only once are not likely to be useful).

These will tend to consist of pre-existing phrases and idioms. Viewing these word clusters is also useful prior to carrying out a concordance.

If numbers are to be excluded, **WordList** will show any numbers as # in the listing. (If you press Ctrl-C for a concordance on such an entry, the # will be replaced by an asterisk in the search word.)

Memory implications

This process uses up RAM at a much faster rate than single-word listing does, because the probability of a cluster being repeated decreases with its length. (A 10 million word corpus might produce 100,000 single types but if you're doing 3-word clusters there would be over a million, which would exhaust your machine's capacity -- and yours in terms of waiting.)

Accordingly, **WordList** may pause, when memory is very low, while it carries out some "house-keeping" clear-up jobs. These involve getting rid of any clusters which have occurred only once so far. Whether this occurs will depend on a) how many texts you're processing, b) how much RAM (not hard disk space) you have in the computer, c) how large the clusters are.

See also: general information on clusters

Detailed Statistics (Σ)

They include:

number of files involved in the word-list
file size (in bytes, i.e. characters)
running words in the text (*tokens*)
no. of different words (*types*)
type/token ratios
no. of sentences in the text
mean sentence length (in words)
standard deviation of sentence length (in words)
no. of paragraphs in the text
mean paragraph length (in words)
standard deviation of paragraph length (in words)
no. of headings in the text
mean heading length (in words)
standard deviation of heading length (in words)
the number of 1-letter words
...
the number of **n**-letter words (to see these scroll the list box down)

(14 is the default maximum word length. But you can set it to any length up to 50 letters in *Word List Settings*, in the *Settings* menu.) Longer words are cut short but this is indicated with a + at the end of the word.

The number of types (different words) is computed separately for each text. Therefore if you have done a single word-list involving more than one text, summing the number of types for each text will not give the same total as the number of types over the whole collection.

See also : Summary Statistics

Summary Statistics

This provides a summary of

Running words (tokens)
Different words (types)
Entries currently in list
Lemmatised entries
Entries in lemmas
Entries matching user-list
Tokens in the entries matching user-list

and ratios of each one of these to the other, where applicable.

Running words (tokens)

This is the number of all the running words encountered as **Wordlist** processed the text file(s). A word is counted as a string of valid letters with a separator at each end. Hyphenation decisions in *Settings* affect whether a string like *self-help* counts as one word or two.

Different words (types)

These are the different words stored as **Wordlist** first processed the text. A file containing 1 million running words might have 50,000 different words. They would include different forms of the same lemma, e.g. *swim*, *swimming*, *swam*, *swims*. Only words meeting your settings criteria (e.g. minimum and maximum length, hyphenation, numbers) are included in this number.

Entries currently in list

As you edit a list, you might decide to delete some items which you did not wish to retain, or correct any mis-spellings, or lemmatise (join *swimming*, *swam*, *swims* to *swim*). The number of entries in your list is therefore likely to shrink.

Lemmatised entries

This is the number of entries which have been lemmatised (*swim* in our example would count as a lemmatised entry).

Entries in lemmas

This is the number of variants of each lemma. In our example, *swimming*, *swam*, *swims* add 3 to this number.

Entries matching user-list

You may have chosen to relate your word list to a file containing certain specific words which you are especially interested in, e.g. to compute a lexical density score. If so, the number here will show how many of the entries in your word list match those in the file. Matched words are marked with a tilde (~).

Tokens in the entries matching user-list

The item above shows how many of the entries in your word list match those in the file. This one shows the total number of tokens in these entries.

Ratios

As long as both of any pair of the above numbers is greater than 0, the ratio between them is presented in the scrollable listbox, as a percentage. The percentage is always of the smaller in relation to the larger. The conventional type/token ratio (A:B in the display) is computed by the formula:

(number of types x 100) ÷ number of tokens

See also : Detailed Statistics

Type/Token Ratios and the Standardised Type/Token ratio

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 ratio 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 type/token ratio 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 and can be re-computed using the Summary Summary Statistics button ().

Wordlist uses a different strategy for computing this, therefore. The *standardised type/token ratio* 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.)

Adjust the ***n*** number in Minimum & Maximum Settings to any number between 10 and 20,000.

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

The number shown is a percentage of new types for every ***n*** tokens. That way you can compare type/token ratios across texts of differing lengths. This method contrasts with that of Tuldava (1995:131-50) who relies on a notion of 3 stages of accumulation.

Joining Entries Together (Lemmatisation)

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

manual joining

To do this you

1. Use F5 to mark an entry for joining to another. The first one you mark will be the "head". For the moment, while you're still deciding which other entries belong with it, the whole row will be marked in a certain colour. Any entries which you then decide to link with the head (by again pressing F5) will show they're marked in a different colour.

If you change your mind you can press F5 again and the marking will disappear.

2. Use F4 to join all the entries which you've marked. The program will then put the joint frequencies of all the words you've marked with the frequency of the one you marked first (the head). Both the alphabetical and the frequency lists will be correctly updated, though the frequency list may not reflect the true order until after the file has been re-ordered by zapping entries. A lemmatised head entry has **...** beside it. The others you marked will be coloured grey. The linked entries which have been joined to the head can be seen by clicking on the head with the right mouse button.

file-based joining

Alternatively you can join up lemmas using a text file which automates the matching & joining process.

auto-joining

To speed up this lemmatisation process, you can use the menu to search for an entry, and if you use the ***** symbol in your search string, the program will seek the first item which matches your lemmatisation criterion. Thus

> ***S** will search for the next *consecutive pair* of entries, where one differs from the other only by the presence of S at the end of the word.
>
> ***ED** would find all entries ending in **-ED**.

The pair will be marked for lemmatisation, so if you agree that they belong together, press F4. If not, press F5 and the marking will disappear. Finally, press F8, to repeat the search and so proceed to the end of the word-list.

This routine works by pairs of consecutive entries. To lemmatise a set such as

> **HELP**
> **HELPED**
> **HELPFUL**
> **HELPING**
> **HELPS**

specify ***ED;S;ING;FUL** as your search string -- you can specify the alternatives, separated by semi-colons, in any order.

To undo a lemma, place the highlight on the head entry and press F4. If the associated entries have not yet been deleted, they will be revived and frequencies adjusted accordingly.

If you save the file or zap any deleted entries, the greyed-out members of the lemmas will disappear from the listing (though they are still viewable by right-clicking on any head entry with **...** beside it).

See also: Auto-Join

Auto-Joining Members of a Lemma

The menu option *Auto-Join* can be used to specify a string such as ***S** or ***S;ED** and will then go through the whole word list, lemmatising all entries where one word differs from the next by having **S** or **ED** on the end of it. At the end, you'll be notified as to how many entries were lemmatised.

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 undo Auto-Joined lemmatisations, just run Auto-Join again with the same specifications.

To stop in the middle of auto-joining, press **Esc**ape.

See also: Lemmatisation

Lemma Match ≡

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.). To do this, press the *Lemma Match* button ≡. 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 less "hit-and-miss" than auto-joining.

How to do it

A dialogue box appears, asking for the file name: type in the name of a file containing a plain text list of lemmas with items like this:

be -> am,are,was,is,were
go -> goes,going,went

WordList then reads the file and checks every entry in your current word list to see whether it matches one of the entries in your text file. In the example, if, say, ***am, was***, and ***were*** are found, they will be stored as lemmas of ***be***. If ***goes*** and ***went*** are found but not ***go***, then ***went*** will be joined to ***goes***.

See also: Lemmatisation, Match List

Search using the Menu

Using the menu you can search for a sub-string within an entry -- e.g. all words containing "fore" (by entering ***fore*** -- the asterisk means that the item can be found in the middle of a word, so ***fore** will find *before* but not *beforehand*, while ***fore*** will find them both). These searches can be repeated.

This function enables you to find parts of words so that you can edit your wordlist, e.g. by joining two words as one.

You can search for ends or middles of words by using the * wildcard.

Thus ***TH*** will find *other, something*, etc.

***TH** will find *booth, sooth*, etc.

You can then use **F8** to repeat your last search.

The search hot keys are:

F8 repeat last search (use in conjunction with F10 or F11)
F10 search forwards from the current line
F11 search backwards from the current line
F12 search starting from the beginning

This function is handy for lemmatization (joining words which belong under one entry, such as *seem/ seems/ seemed/ seeming* etc.)

See also: searching for an entry by typing

WordList Sorting (or F6)

Many languages have their own special sorting order, so prior to sorting or re-sorting, check that you have selected the right language for the words being sorted. Spanish, for example, uses this order: **A,B,C,CH,D,E,F,G,H,I,J,K,L,LL,M,N,Ñ,O,P,Q,R,S,T,U,V,W,X,Y,Z**

À with A, Ç with C

Accented characters are by default treated as equivalent to their unaccented counterparts in some languages (so, in French we get **donné, donner, donnes, 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** before **čas**)

If this box is checked, sorting will treat accented characters as related to their unaccented counterparts according to the current language.

Reverse Sort

This is so that you can sort words by suffix. The order is determined by word endings, not word beginnings. You will therefore find all the *-ing* forms together.

See also: Editing entries; Accented characters; Choosing Language

Comparing Wordlists: the purpose

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.

The minimum frequency (which you can alter in the Controller, *Adjust Settings, KeyWords* tab) can be set to 1. If it is raised to say 3, the comparison will ignore words which do not appear at least 3 times in at least one of the two lists.

Choose the significance value (*all*, or a p value from 0.1 to 0.000001 or what you will). The smaller the p value, the more selective the comparison. In other words, a p setting of 0.1 will show more words than a p setting of 0.0001 will.

To compare two word-lists, simply choose the two lists (Word List 1 and Word List 2).

The display format is similar to that used in KeyWords.

See also: Consistency Analysis, Match List

Comparing Word Lists: the display

The display shows

frequency in smaller text (with % if > 0.01%) -- then, to the right

frequency in larger text (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*. At the end of the listing you'll find those which are outstandingly *infrequent*.

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.

Consistency Analysis (Simple)

This function compares word lists, and allows many more than 50 word lists to be analysed.

With this procedure you can process a large number of texts and produce a word list which shows the frequency in terms of the number of texts (as opposed to frequency in terms of the running words in each text).

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 new consistency word list, just like any other, 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

Consistency Analysis (Detailed)

This function does exactly the same thing as simple consistency, but can provide more detail, as long as no more than 50 existing word lists are chosen for analysis. (If you need more, go for simple consistency analysis.)

The point of it…

The idea is to help stylistic comparisons. Suppose you're studying several versions of a story, or different translations of it. This function enables you to see *all* the words which are used in the wordlists which you have called up. The display will order the words, so that the first group contains all those which occur in all versions, then those which come in all versions but one, and so on down to those which occur in only one version.

Example: 3 translations of *Candide* might show:

...

garden	**3**	5	**1**	1
			7	1
...				
Pangloss	**3**	**3**	3	3
		7	5	6
...				
bridge	**2**	0		
			3	**4**

...

Within each set the words are ordered alphabetically. The left column (in blue here) shows the total number of text files where the word appeared. Subsequent columns give the number of occurrences for each text file.

You can also see how frequent each one was in each of the texts. The highest frequency will appear in red.

In this example, the second translator used *garden* much more than the others, who must have used other translations such as anaphoric *it* or *there*. One translator did not use the term *bridge* at all.

See also: Consistency Analysis (Simple), Comparison Display, Comparing Word-lists, Match List

Re-sorting Consistency Lists (or F6)

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.

Splitter Index

Explanations

What is the Splitter program and what's it for?
Filenames
Wildcards
see also : WordSmith Main Index

What is *Splitter* and what's it for?

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

End of Text Separator

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

Restrictions:

1 The end-of-text marker must occur at the beginning of a line in the original large file.

2 It is case sensitive: **</Text>** will not find **</text>**.

3 The first character in the end-of-text separator may not be a wildcard such as **#,*** or **?**.

4 ***** and **#** may occur only once each in the end-of-text separator.

Destination Directory

where you want the small files to be copied to. (You'll need write permission to access it if on a network.)

Required sizes

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

A "line" means from one <Enter> to the next, up to a maximum of 10,000 characters.

<bracket first line>

whether or not you want the first line of each new text file to be bracketed inside < > marks. This is because often the first line after your end-of-text symbol will contain some kind of header. If you don't want it to insert < and > around the line, leave the checkbox un-checked.

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

see also: Filenames, Wildcards, The buttons, **Text Converter**.

Wildcards in Splitter

- # The hash symbol, #, is used as a wildcard to represent any ***number***, so **[FF#]** would find [FF3] or [FF9987] but not [FF] or [FF 9] (because there's a space in it) or [FFhello].
- * The asterisk represents any ***string***, so **[FF*** would find all of the above. * is used as the last character in the end-of-text symbol. It would find *[FF anything at all up to the next <Enter>*.
- **?** The question mark represents any single ***character*** (including spaces, punctuation, letters), so **[FF??]** would find [FF 9] in the above examples, but none of the others.

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

see also: Settings, Wildcards

Filenames made by Splitter

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

It creates new filenames on the following basis:

The first two letters of each large file are retained, with a further 6-digit number. **txt** is added as a file extension.

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

he000001.txt
he000002.txt
etc.

Tips

1. It's best to create a new directory for the small text files before proceeding. To do this, use File Manager.
2. **Splitter** will start numbering at 1 each session.
3. 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 Wordlist Batches or KeyWords Batches.
4. CD-ROM files when copied to your hard disk will be read-only. You can change this attribute

using Text Converter.

Text Converter Index

Explanations

What is *Text Converter* and what's it for?

This program does four jobs, on up to 16,368 files.

renaming files

If you have extracted lots of text files from large CD-ROM texts, e.g. using **Splitter**, you may well find the filenames are inconvenient and wish to rename large numbers in one fell swoop. Click here to see how.

changing attributes

Similarly, CD-ROM text files, as their name indicates, are read-only. To edit them you will need to change their "attributes". Click here to see how.

Moving files "if"…

This function looks at your text files and can move them into a new directory if they contain certain words or phrases. Click here to see how.

converting text

For a simple search-and-replace you can type in the search item and a replacement; for more complex conversions, use a Conversion File so that **Text Converter** knows which symbols or strings to convert. It operates under Windows and saves using the Windows character set, but will convert text using DOS or Windows character sets. You can use it to make your text files suitable for use with an Internet browser such as Netscape.

It does a "search and replace" much as in word-processors, but it can do this on up to 16,368 text files, one after the other. As it does so, it can also replace up to **500** 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, ensuring you have Windows **£** symbols, etc.

Once the conversion file is prepared and Settings specified, the **Text Converter** will read each source file and either create a new version or replace the old one, depending on the over-write setting.

You will be able to see the details of how many instances of each string were found and replaced overall.

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 directory and try out your conversion file with the Text Converter. You may find you've failed to specify some necessary conversions. Once you're sure everything is the way you want it, delete the temporary files.

See also: Text Converter Contents, The buttons

Text Converter: Getting Started…

1. Choose text files in the top left box. Decide whether you want the program to process sub-directories 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 Activated box for Text Conversion.
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. (You will be warned if disk space goes below 1 megabyte).
4. If you choose to make copies, specify where to put the results, that is the destination directory. This must be different from your source directory.

   Example and default: **c:\temp**. If necessary the Text Converter will create the directory you specify as long as you specify a legal name. If you have chosen to process sub-directories too, the same directory structure will be created, stemming from this directory.
5. Specify what to convert, that is the search-words and what you want them to be replaced with. For a quick conversion you can simply type in a word you want to change and its replacement (e.g. Just one change so that **responsable** becomes **responsible**) or you can type a filename and directory of the Conversion File e.g. **c:\text\convert.sym.**

If you choose *Over-write Source texts,* Text Converter will work more quickly and use less disk space, but of course you should be quite sure your conversion file codes are right before starting!

Note that ***some space on your hard disk will be used even if you plan to over-write***. The conversion process does its work, then if all is well the original file is deleted, and the new version copied. There has to be enough room in the destination directory for the largest of your new files; it is much quicker for it to be on the same drive as the source texts. If it isn't, your permission will be asked to use the same drive.

Press *Go Now* to start; you can review results with the *Results* button.

If you want to stop **Text Converter** at any time, click on the Cancel button or press Escape.

See also Text Converter Contents.

Changing Attributes

CD-ROM text files are read-only and the ones copied from CD-ROMS onto your hard disk will also be read-only. To edit them you will need to set the read-only "attribute" to un-checked.

The attributes

archive: an ordinary text file
read-only: if checked, cannot be edited
hidden: if checked, may not get listed in some displays (best avoid checking this one!)
system: if checked, may have special significance in the system (avoid this one, too!)

Choose the combination you want, after first clicking on the Activated checkbox.

See also Text Converter Contents.

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 directory.

The point of it …

Suppose you have a whole set of files some of which contain dialogues between Pip and Magwich, others containing references to the Great Wall of China or the anatomy of fleas. You want those with the Pip-Magwich dialogues and you want them to go into a directory called *Expect*.

How to do it

First click on the Activated checkbox.

Then decide how many lines you wish to examine in each text file (default = 10). The more lines, the longer it'll take, naturally. A line is up to an <Enter> or 10,000 characters.

Now specify a word or phrase the text must contain. This is case sensitive.

Finally type in the directory name. You would specify "*expect*" in this space.

This is how the directory system works in this function: suppose your texts are in c:\texts, but you're searching through all the sub-directories too. You will therefore have left the top left directory listing and the label above "process sub-directories" showing **c:\texts**. Now suppose a file containing *Magwich* is found in a directory called c:\texts\a\xxx. In that case it will be placed in c:\texts\expect\a\xxx. If another file containing *Magwich* is found in a directory called c:\texts\ppp, it will be placed in c:\texts\expect\ppp. In other words, what you specify here provides a "branch" off which new branches can be created, and the branch names will reflect the original directory structure. The sub-directories will be created automatically, by the way.

See also Text Converter Contents.

Renaming Files

The format is still 8.3 – that is, up to eight characters, a dot, up to three characters. Acceptable characters are A to Z, _ and 0 to 9.

The "mask" means a template for the changes you want.

examples

***.123** forces all files to have .123 as the file extension.

LIST????.TXT will change all files so that the first four letters are LIST, then up to 4 letters of the original name will be used, then .TXT will be forcibly used as the file extension.

L*.TXT will change all files so that the first letter is L, then up to 7 letters of the original name will be used, then .TXT will be forcibly used as the file extension.

L####.* will change all files so that the first letter is L, followed by a new 4-digit number and the original file extension.

When the name is too restrictive so that another file already exists with that same name, you will be asked whether to abort the operation, or allow a new name to be created with a number in it, so that the name is unique.

Type in the mask, after first clicking on the Activated checkbox.

See also Text Converter Contents.

Text Converter Syntax

The syntax for a Conversion File is:

Only lines beginning / or **"** are used. Others are ignored completely.

Every string for conversion is of the form **"A" -> "B"**. That is, the original string, the one you're searching for, enclosed in double quotes, is followed by a space, a hyphen, the > symbol, and the replacement string.

Control Codes

Control codes can be symbolised like this: {CHR(xxx)} where xxx is the number of the code. Examples: {CHR(13)} is a carriage-return, {CHR(10)} is a line-feed, {CHR(9)} is a tab, {CHR(12)} is a printer form-feed. To represent <Enter> which comes at the end of paragraphs and sometimes at the end of each line, you'd type **{CHR(13)}{CHR(10)}** which is carriage-return followed immediately by line-feed.

Use {CHR(34)} if you need to refer to double inverted commas.

Wildcards (*,?,# and ~)

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

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

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

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

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

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

Whole word, case Insensitive, Confirm, First part only, redundant Spaces

/C stops to **c**onfirm you wish to go ahead before each change.

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

/I does a case **i**nsensitive search (/I "restaurant" -> "hotel" replaces **restaurant** with **hotel** and **RESTAURANT** with **HOTEL** and **Restaurant** with **Hotel**, i.e. respecting case as far as possible).

/F searches & replaces only in the **f**irst part of each text file (within the first 30,000 characters, approx. 5,000 words). This is somewhat speedier in the case of large text files.

You can combine these, e.g.

/IWC "the" -> "this"

/S " " cuts out all redundant spaces. That is, it will reduce any sequence of two or more spaces to one, and it also removes some common formatting problems such as a lone space after a carriage-return or before punctuation marks such as **.,;** and **)**. **/S "{CHR(9)}"** will likewise cut out all redundant tabs, reducing any sequence of two or more tabs to one. **/S** should be used on a line of its own.

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

/A means add text. **/A "Ulan" START** inserts *Ulan* at the start, **/A "Bator" END** inserts *Bator* at the end of the text. See convert.txt (in your \wsmith directory) to see one in use.

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

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

See convert.txt (in your \wsmith directory) to see examples in use.

See also: Text Converter Contents.

Text Converter Conversion File

Prepare your Text Converter conversion file using a plain text or Windows editor.

If the files are in English, it makes little difference whether you use for example **edit.com** (DOS) or **notepad** (Windows). Click here to see stoplist.cod or convert.txt (in your \wsmith directory) in **notepad** which you could use as a basis.

If you have accented characters in your original files, use the DOS editor to prepare the conversion file if they were originally written under DOS and a Windows editor if they were written in a Windows word-processor. Some Windows word processors can handle either format.

There can be up to 500 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.

Alterations that *increase* the original file

Most changes reduce the size of an original. But if you're making changes that *increase* it, e.g. replacing "***the***" by "***much more than just the***", there will be a slight danger of over-flow. The space allocated in the program is exactly **double** the space used by the original text (60,000 characters). There will only be an overflow if you are inserting a lot of extra characters very frequently (and you'll be told if one happens). In this example, overflow will occur if the 3-letter string ***the***, to be replaced by the 23 letters of ***much more than just the*** crops up more often than every 20 characters in the original text. In practice this is extremely unlikely to happen unless you are making a whole series of alterations that each increase file size. In that case, do the conversions in separate runs, using different Conversion Files on each run.

Tips

1. To get rid of the <Enter> at line ends but not at paragraph ends, first examine your paragraph ends to see what is unique about them. If for example, paragraphs end with two <Enters>, use the following lines in your conversion file:

"{CHR(13)}{CHR(10)}{CHR(13)}{CHR(10)}" -> "%%^"

(this line replaces the two <Enters> with %%^ (it could be any other unique combination)

"{CHR(13)}{CHR(10)}" -> " "

(this line replaces all other <Enters> with a space, to keep words separate)

"%%^" -> "{CHR(13)}{CHR(10)}{CHR(13)}{CHR(10)}"

(this line replaces the %%^ combination with <Enter><Enter>, thus restoring the original paragraph structure)

"/S " "

(this line cuts out all redundant spaces)

2. The file stoplist.cod which came with your installation (in your \wsmith directory) will format a word list if *Saved as Text* and make a stop list from it.

See also: sample conversion file, syntax, Text Converter Contents.

Sample Text Converter 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"

[now let's change brackets from < > to [ ] and { } to ( )]

"<" -> "["

">" -> "]"

"}" -> ")"

"{" -> ")"

/S " "

[that will clear all redundant spaces]

The file convert.txt (in your \wsmith directory) is a sample conversion file for use with British National Corpus text files.

See also: Text Converter Contents.

Viewer Index

Explanations

What is *Viewer* and what's it for?

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

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

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

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

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

See also: Viewer settings, Viewer options

Viewer Settings

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

Colours
Font
Printing

Text Characteristics
Review all Settings

Different Views of your texts: the View menu

Mode: Sentence/Paragraph

This switches between Sentence mode (the default) and Paragraph mode. In other words you can choose to view your text files with each row of the display taking up a sentence or a paragraph. Likewise, you can make an dual aligned text by interspersing either paragraphs or sentences. The other functions (e.g. joining, splitting) work in the same way in either mode.

If you re-save the text, you will probably prefer paragraph mode; this will put an <Enter> at the end of each paragraph. In sentence mode you will get an <Enter> at the end of each sentence, giving a text which is in effect a long list of sentences.

Display: Normal/Accents etc. as Codes

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

Ignored tags: cut/visible

Your Tags To Ignore can be made visible or invisible.

Format: Plain Text/HTML/SGML

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

Header: cut/visible

If your text is tagged and has a header, this too can be shown or cut out.

Editing texts in Viewer

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

Edit (✓): opens up a window allowing you to edit the whole of the current sentence or paragraph.

Trim extra spaces: this goes through each sentence of the text, removing any redundant spaces -- where there are two or more consecutive spaces they will be reduced to one.

Find lower-case lines: this identifies cases where a sentence or paragraph does not start with a capital letter or number -- you will probably want to join it to the one above. This problem is common if the text has been saved as "text only with line breaks" (where an <Enter> comes at the end of each line whether or not it is the end of a paragraph.)

Find short lines

Insert tags

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

You can also:

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

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

print the whole or part of the currently active text file (unless it has more than 16,328 sentences)

change the current character set

search for words or phrases (press F12)

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

Sentence or Paragraph Joining and Splitting

Joining ⁝}

The easiest way to join two sentences is simply to drag the one you want to move onto its neighbour above. Or you can mark them with ●, then press ⁝}. You will be told if you're in danger of jumbling up your text!

Splitting in two ✗

To split a sentence, press ✗. You will get a list of the words. Click on the word which should ***end*** the sentence, then press OK.

example

"It
is
good!"
Mary
wanted
more.

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

See also: Viewer contents

Using Viewer to Tag Sentences and/or Paragraphs

You can use the **Viewer** to make a copy of your text with all the sentences and/or paragraphs tagged with **<S> and <P>**.

To do this, simply read in the text file in, choose *Edit* | *Insert Tags*, then save it as a text file.

To choose between sentence or paragraph mode, see Settings.

See also: Viewer contents

Seeking Unusual Sentences (F8)

Because **Viewer** uses full stops, question marks and exclamation marks as sentence-boundary indicators, you will find a string like "Hello! Paul! Come here!" is broken into 3 very short sentences. Depending on your purposes you may wish to consider these as one sentence, e.g. if a translator has translated them as one ("Oi, Paulo, venha cá!") .

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

This function also finds lower-case lines: where a sentence or paragraph does not start with a capital letter or number -- you will probably want to join it to the one above. This problem is common if the text has been saved as "text only with line breaks" (where an <Enter> comes at the end of each line whether or not it is the end of a paragraph.)

seeking

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

See also: The technical side..., Finding translation mis-matches, Viewer contents

Technical Aspects in Viewer

When is a sentence not a sentence?

There is no perfect mechanical way of determining sentence-breaks. For example, a heading may well have no final full stop but would normally not be considered part of the sentence which follows it. And a sentence may often have no final full stop, if what follows it is a list of items.

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

Consider this chunk from *A Tale of Two Cities*:

"Wo-ho!" said the coachman. "So, then! One more pull and you're at the top and be damned to you, for I have had trouble enough to get you to it! - Joe!"

Viewer will mistakenly consider *- Joe!* as a separate sentence, but handles *"Wo-ho!" said the coachman.* as one: though the program would split it in two if the word after *ho!* had a capital lettter (e.g. in *Wild Bill, the coachman, said.*)

Viewer cannot therefore be expected to handle all sentence boundaries exactly as *you* would. (*I saw Mr. Smith.* would be considered two sentences; several headings may be bundled together as one sentence.) For this reason you can choose *Find Short Sentences* to seek out any odd one-word sentences.

How long is a sentence?

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

Disk handling, Accents, etc.

For best results, use ascii or ansi versions of your two texts.

The joint text will be saved using a Windows character set.

See also: Viewer contents

Viewer Trouble-shooting

Can't see the whole sentence or paragraph

Press ⇕ to "auto-size" the lines in your display. This adjusts line heights according to the current highlighted column of data.

Can't see the whole text file

Press □ to "refresh" the display. Viewer will re-read the text file up to its limit of 16,368 sentences.

Don't like the colours

Change colours in the top-level *Settings* menu. The colours used for each language version in the dual-language window are the same colours as used for primary sorting and secondary sorting in **Concord**.

See also: Viewer contents

What is Aligning for?

This feature aligns the sentences in two files. Translators need to study differences between an original and a translation. Other linguists might want it to study differences between two versions of a text in the same language. Students of different languages can use it as they might use dual language readings, to study closely the differences e.g. in word order.

It helps you produce a new text which consists of the two files, with sentences interspersed. That way you can compare the translation with the original.

Example

Original : *Der Knabe sagte diesen Gedanken dem Schwesterchen, und diese folgte. Allein auch der Weg auf den Hals hinab war nicht zu finden. So klar die Sonne schien, ...(from Stifter's Bergkristall, translated by Harry Steinhauer, in German Stories, Bantam Books 1961)*

Translation: *The boy communicated this thought to his sister and she followed him. But the road down the neck could not be found either. Though the sun shone clearly, ...*

Aligned text:

<G1> Der Knabe sagte diesen Gedanken dem Schwesterchen, und diese folgte.

<E1> The boy communicated this thought to his sister and she followed him.

<G2> Allein auch der Weg auf den Hals hinab war nicht zu finden.

<E2> But the road down the neck could not be found either.

<G3> So klar die Sonne schien, ...

<E3> Though the sun shone clearly, ...

An aligned text like this helps you identify additions and omissions, normalisations, style changes, word order preferences. In this case the translator has chosen to avoid very close equivalence.

See also: The buttons

Aligning the Dual Text

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

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

adjusting by dragging with the mouse

To alter sentence order, just drag a sentence to the next one above of the same colour.

Finally you will want to save (F2) the results.

See also: Viewer contents

Seeking Translation Mis-matches

Viewer can help find cases where alignment has slipped (one sentence having been translated as two or three). One method is to use the menu item *Match by Capitals*. This searches for matching proper nouns in the two versions: if say Paris is mentioned in sentences 25 of the source text and not in sentence 25 of the translation but in sentence 27, it is very likely that some slippage has occurred.

Viewer will search forwards from the current text sentence on, and will tell you where there's a mis-match. You should then search back from that point to find where the sentences start to diverge. It may be useful to sample every 10 or every 20 to speed up the search for slippage.

When you find the problem, un-join or join and/or edit the text as appropriate, then save it.

See also: The technical side..., Finding unusual sentences, Viewer contents

Some Definitions

concordance: a set of examples of a given word or phrase, showing the context. A concordance of

give might look like this:

```
... could not give me the time ...
... Rosemary, give me another ...
... would not give much for that ...
```

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

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

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

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

The DOS program **edit.com** makes plain ASCII text files. The Windows program **notepad** makes so-called ANSI text files, sometimes referred to as **.txt** files. These formats use the same character sets for the English alphabet from A to Z, numbers and common punctuation symbols. The main difference is in the accented characters. For more on this, see character sets.

See also : HTML,SGML & XML.

HTML, SGML, 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

long file names. Windows 3.x uses short filenames (no spaces, eight characters, a dot, three characters, or 8.3 format) whereas Windows 95 and above allow long filenames with up to 255 characters, spaces, etc.

If you're running **WordSmith** in Windows 95, and you choose a file by browsing, as opposed to typing it in, any long names will automatically be translated into 8.3 format by Windows. You should be aware that though this translation works, the abbreviated name might come out differently the next time. The safest method of ensuring that your filenames will be satisfactory is to use 8.3 format and to avoid creating or using directories with long names or spaces in them (e.g. ***\Program Files\Internet Explorer***). If you need to refer to ***c:\Program Files\Internet Explorer\iexplore.exe***, I suggest you create a shortcut to it in a directory with ordinary 8.3 format name, such as ***c:\shorts\*** then rename the "Shortcut to iexplore.exe.lnk" as ***iexplor.lnk.*** You can use the ***\shorts*** directory as a general device for getting round this problem with 16-bit programs.

Definitions

words

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

A word can be of any length but for one to be stored in a word list, you may set the length you prefer (maximum of 50 characters) -- any which exceed your limit will get + tagged onto them at that point. You can decide whether or not to include words including numbers (e.g. $35.50) in text characteristics.

clusters

A cluster is a *group of words which follow each other in a text*. The term *phrase* is not used here because it has technical senses in linguistics which would imply a grammatical relation between the words in it. In WordList cluster processing or Concord cluster processing there can be no certainty of this, though clusters often do match phrases or idioms. See also: general cluster information.

sentences

The sentence is defined as *the full-stop, question-mark or exclamation-mark (.?!) immediately followed by one or more word separators and then a capital letter A..Z or an accented capital letter, a number or a currency symbol*. (For more discussion see Viewer technical information.)

paragraphs

Paragraphs are user-defined. See Tagged Text for further details.

headings

Headings are also user-defined.

See also: Setting Text Characteristics, Key-ness, Key key-word, Associate

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.

Trouble-Shooting

When it doesn't do what you expected...

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

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.

(Click here to see them in Character Map. Choose Arial font and look below q,r,s and t.)

Solution: drag the symbol from the set below when entering your search word, or else replace them in your text files using Text Converter.

See also: settings

weird symbols

funny symbols when using WordSmith Tools

1. Check your text files. Read them in **Notepad**. Do they contain lots of strange symbols? These may be hidden codes used by your usual word-processor. Solution: read them into your usual word-processor and *Save As*, with a new name, in plain text format, sometimes called "Text Only" or **.txt**.

2. *Choose Texts*, highlight the text file, and before pressing *OK*, press *View*. Does it contain strange symbols? Solution: change *Text Settings*; try going from one of the DOS character sets to Windows

or vice-versa. The text was clean ASCII but **WordSmith Tools** thought it was Windows ANSI.

3. Funny symbols in a word list may well also be caused by mis-spellings in the original text files.

Greek, Russian, etc.

4. If the text is in Russian, Greek, etc. you will need an appropriate font, obtainable e.g. for Windows 95 via Microsoft Plus.

5. If you have several lists open which use *different* character sets, and you change Font or Text Characteristics, the lists will all be updated to show the current font and character set, unless you first minimize any window which would be affected.

funny symbols when reading WordSmith data in another application

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 Tools**. This format is not suitable for reading into a word processor. The idea is simply for you to store your work so that you can return to it another day.

"Save as Text", on the other hand, means saving as plain text, by "printing" to a file. This function is useful if you don't want to print to paper from **WordSmith** but instead take the data into a spreadsheet, or word processor such as **WordPerfect** or **Microsoft Word**. It is usually quicker to copy the selected text into the clipboard.

Concord/WordList mismatch

If **WordList** finds a certain number of occurrences of a (word list) cluster but **Concord** finds a different number, this is because the procedures are different. WordList proceeds word by word, ignoring punctuation (except for hyphens and apostrophes). When **Concord** searches for a (concordance) cluster it will take punctuation into account.

Keys don't respond

If a key press does nothing, it is probably because the wrong window has the focus. As you know, Windows is designed to let users open up a number of programs at once on the same screen, so each window will respond to different key-press combinations. You can see which window has the focus because its caption is coloured differently from all the others. The solution is to click anywhere within the window which you want to use, then press the key you wanted.

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.

printing problem

If your printing comes out with one or more column blank but others printed correctly, you may have a printer which can only manage black and white and not shades of grey. In the Controller, change the setting (*Adjust Settings* | *General*) to monochrome.

If you have difficulty printing on a network in Windows 95, check whether the network administrator has enabled printing from DOS programs, which may be necessary. Thanks to Donald MacQueen, Dept. of English, Uppsala University for this information.

column spacing is wrong

You can alter this by clicking on the layout button.

demo limit reached

You may have installed (with **setup.exe**) and typed in your name, but you haven't typed in the registration code. To do this, go to the main WordSmith Tools window, and choose *Demo* | *Update from Demo Version* in the menu.

If you haven't got the 20-character registration code, contact Oxford University Press (Click here (in your \wsmith directory) for details). 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.

it doesn't even start

You have to install by running **setup.exe** from your floppy disk. You also need to be running Windows 3.1 or greater, and enough memory to run **WordSmith Tools**: at least 2MB of RAM and preferably much more!

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.

It takes ages

If you're processing a lot of text and you have an ancient 386 with only 2 MB of 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: read from the hard disk, not the CD-ROM.

Solution 5: use Windows 3.1 instead of Windows 95: it'll give you about 2MB more RAM.

it crashed!

Solution: quit **WordSmith Tools** and enter again. If that fails, quit Windows and try again.

wordlist out of order

In the case of accented characters, any accented letter will immediately follow the same letter without the accent. Thus, by default the French word ÉTÉ will (wrongly) follow EZ. In many cases this sort order will be adequate. The worst cases come where the first letter of a word is accented; if the fifth letter is accented the imperfection in the order is unlikely to be noticeable.

Solution: press ⟳ to re-sort a word list, if you want words such as ÉTÉ or ÉTERNEL to come between ET and ETHNIQUE. (Je l'ai fait comme ça pour accélérer la classification pour la plupart des gens: à vous de re-classer si vous voulez! Ou seja, você sempre pode botar em ordem depois se assim quiser... para tomar en cuenta la distinción entre la L y la LL, or in other languages get a special alphabetical order, re-sort.)

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.

Error Messages

List of Error Messages

See also: Troubleshooting.

Can only save WORDS as ASCII

Can't call other Tool

Can't make directory as that's an existing filename

Can't merge list

Can't read file

Can't show all files in directory (though they are sorted correctly)

Character set reset to <x> to suit <language>

Concordance file is faulty

Concordance stop list file not found

Conversion file not found

Destination directory not found

Disk problem: File not saved

Dispersions go with concordances

Drive not valid

Failed to access Internet

Failed to create new directory name

File access denied

File contains none of the tags specified

File not found

Filenames must differ!

Full drive:\directory 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

Limit of 20 windows reached

Links between Tools disrupted

Match list details not specified

Must be a number
Network registration running elsewhere or vice-versa
No access to text file: in use elsewhere?
No associates found
No clumps identified
No clusters found
No collocates found
No concordance entries found
No concordance stop list words
No deleted lines to Zap
No entries in Keywords Database
No Key Words found
No key words to plot
No keyword stop list words
No lemma list words
No match list words
No room for computed variable
No statistics available
No stop list words
No such file(s) found
No tag list words
Not a valid number
No wordlists selected
Original text file needed but not found
Registration string is not correct
Registration string must be 20 letters long
Run SETUP on distribution disk to install
Short of Memory!
Source Directory 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 3.1 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

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.

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 **wshell.exe** program again.

Can't make directory as that's an existing filename

If you already have a *file* called C:\TEMP\FRED, you can't make a *sub-directory* of C:\TEMP called FRED. Choose a new name.

Can't merge list

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.

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, 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.

Can't show all files in directory (though they are sorted correctly)

Under Windows 3.1 there is a limit to the number of files which can be shown in the Choose Files list. Files will be sorted (by name, date or size) according to your choice, but you will only be able to see part of the details of each file.

Windows 95 will show all your file name information (memory permitting).

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, *Adjust Settings* | *Text*, then re-save the list.

Concordance file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

Concordance stop list file not found

You typed in the name of a non-existent file. If typing in a filename, remember to include the full drive and directory as well as the filename itself.

Conversion file not found

You typed in the name of a non-existent file. If typing in a filename, remember to include the full drive and directory as well as the filename itself.

Destination directory not found

WordSmith couldn't find that directory; perhaps it's mis-spelt.

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.

Dispersions go with concordances

They can't be saved separately.

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:.

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, c) the browser's filename being recognised by Windows. If you get this error message, ensure the full filename of your Internet browser (e.g. **c:\netscape\netscape.exe**) is specified in *Adjust Settings* | *General*.

Failed to create new directory

A directory and a file cannot have the same name. If you already have a *file* called C:\TEMP\FRED, you can't make a *sub-directory* of C:\TEMP called FRED. Choose a new name.

Failed to Read

This may have happened because your disk filing system has got screwed up. This is especially likely to occur if you often use large files in your word processor. I would recommend you to run *System Tools* | Scandisk.

Failed to Save

This may have happened because the directory you're saving to is a read-only directory on a network, or because the disk is full, or because your disk filing system has got screwed up. This last problem is quite common, actually, and is especially likely to occur if you often use large files in your word processor. I would recommend you to run *System Tools* | Scandisk. If you're working on a network, you will be able to save on certain drives and directories but not others; the solution is to try again on drive A: or a hard disk drive which you have the right to save to.

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.

File contains none of the tags specified

You specified tags, but none of them were found.

File not found

This message, like Original Text not found, can appear when **WordSmith** needs to access the original source text used when a list was created, but cannot find it. Have you deleted or moved it? If the file is still available, you may be able to edit the filenames in the filename window (▤) of this list.

Or the message may come after you've supplied the filename yourself. You may have mis-typed it. Is it a Windows 95 or NT long filename?. If typing in a filename, remember to include the full drive and directory as well as the filename itself.

Filenames must differ

You can't compare a file with itself.

Full drive:\directory name needed

When typing in a filename, remember to include the full drive and directory as well as the filename itself.

Form incomplete

You tried to close a form where one or more of the blanks needed to be filled in before **WordSmith** could proceed.

function not working properly yet

This is a function under development, still not fully implemented.

.ini file not found

On starting up, **WordSmith** looks for the **wshell.ini** file which holds your current defaults. If you've removed or renamed it, restore it. This file should be in the same directory as the Tools are in.

Invalid Concordance file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

Invalid file name

Filenames may not contain spaces or certain symbols such as **?** and *****. In Windows before Windows 95 they had to be restricted to 8 letters and a dot and three more, too. Try again.

Invalid Keywords Database file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

Invalid Keywords file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

Invalid Wordlist Comparison file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

Invalid Wordlist file

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

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.

Key words file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

Keywords Database file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

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.

Limit of windows reached

This procedure has too many sub-windows open. Close some down and try again.

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.

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.

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).

Network registration running elsewhere or vice-versa

The registration for use on a network is not valid for use on a stand-alone pc, and vice-versa. If you get this message, please re-install as appropriate.

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.

No associates found

Alter settings (*Settings* | *Min & Max Frequencies*) and try again.

No clumps identified

Alter settings and try again.

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.

No collocates found

In the Controller, alter the settings (*Adjust Settings* | *Concord* | *Min. Frequency*) and try again. There were too few concordance lines to find the minimum number needed.

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 whether you've spelt the search-word and context word, specified any wildcards (* and ?) accurately. If you're using accented text, check the format of your texts. If you're using a search-word file, ensure this was prepared using a plain Windows word-processor such as **Notepad**.

Tip

Bung in an asterisk or two. You're more likely to find *book** than *book*.

No concordance stop list words

No deleted lines to Zap

You pressed Alt-Z but hadn't any deleted lines to zap. No harm done.

No entries in Keywords Database

Alter settings and try again.

No Key Words found

Alter settings and try again. The minimum frequency is set too high and/or the p value too small for any key words to be detected. For very short texts a minimum frequency of 2 may be needed.

No key words to plot

Had you deleted them all?

No keyword stop list words

WordSmith either failed to read your stop-list file or it was empty.

No lemma match list words

WordSmith either failed to read your lemma list file or it was empty.

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!

No room for computed variable

There isn't enough space for the variable you're trying to compute.

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.

No stop list words

WordSmith either failed to read your stop-list file or it was empty.

No such file(s) found

You typed in the name of a non-existent file. If typing in a filename, remember to include the full drive and directory as well as the filename itself.

No tag list words

WordSmith either failed to read your tag file or it was empty.

Nothing activated

Some forms have choices labelled "Activated" which you can switch on and off. If they are un-checked, you can still see what they would be but **WordSmith** will ignore them.

No word lists selected

For **WordSmith** to know which word lists to compare, you need to select them, by clicking on one in each directory. If you've changed your mind, press Cancel.

Not a valid number

Either you've just typed in, or else **WordSmith Tools** has just attempted to read (e.g. from **wshell.ini**, the defaults file), something which is expected to be a number but wasn't. Computers will not see the capital **O** as equivalent to the number **0**. Or else there is a number but accompanied by some other letters or symbols, e.g. ***£30***. If this happens when **WordSmith** is starting up, check out the **wshell.ini** file for mistakes.

Okay to re-read?

A confirmation message. To proceed, **Viewer** will now re-read the disk file. This will affect any alterations you've already made to the display. You may wish to save first and then try again later.

Also, Viewer will try to read the whole text file (up to the limit of 16,368 sentences). If you have a very big file on a slow CD-ROM drive, this will take some time.

Original text file(s) needed but not found

To proceed, **WordSmith** needed to find the original text file which the list was based on. But it has been moved or renamed.

Or else the right disk or CD-ROM is not in the drive!

Registration string 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.

Registration string must be 20 letters long

20 letters are needed. You should put them in with a dot after every 4 letters.

Run SETUP to install

WordSmith Tools won't run until you actually install it. The installation file is called **setup.exe**. Double-click on it to start installation.

Short of Memory!

An operation could not be completed because of shortage of RAM

Source Directory file(s) not found

You typed in the name of a non-existent file. If typing in a filename, remember to include the full drive and directory as well as the filename itself.

Stop list file not found

You typed in the name of a non-existent file. If typing in a filename, remember to include the full drive and directory as well as the filename itself.

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.

Tag File not found

You typed in the name of a non-existent file. If typing in a filename, remember to include the full drive and directory as well as the filename itself.

Tag list file not read

Something has gone wrong with a disk reading operation. The file you're trying to read in may be corrupted. This happens easily if you often handle very large files, especially if it's a long time since you last ran Scandisk to check whether any clusters in your files have got lost. See your DOS or

Windows manual for help on fragmentation.

This function is not yet ready!

Temporary message, for functions which are still being tested.

This is a demo version

You will probably want to upgrade to the full version.

This program needs Windows 3.1

It *might* run on Windows 3.0... better to upgrade anyway!

To stop getting this message ...

Get an update. This is "annoyware" for the demonstration version.

Too many requests to ignore matching clumps

The limit is 50. Do any remaining joining manually.

Too many sentences

The limit is 8,000. Do the task in pieces.

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!

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.

Unable to merge Keywords Databases

Perhaps there wasn't enough RAM to carry out the merge.

Why did my search fail?

The standard search function (F12 or 🔍) for a list of data operates on the currently highlighted column. If you want to search within data from another column, click in that column first.

By default, a search is "whole word". Use * at either end of the word or number you're searching for if you want to find it, e.g. in any data consisting of more than one word. (The advantage of the asterisk system is that it allows you to specify either a prefix or a suffix or both, unlike the standard Windows search "whole word" option.)

Word list file not found

You typed in the name of a non-existent file. If typing in a filename, remember to include the full drive and directory as well as the filename itself.

WordList comparison file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

Word list file is faulty

Each type of file created by **WordSmith Tools** has its own default filename extension (e.g. .DOC, .TXT) 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**.

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.

WordSmith Tools already running

Don't try to start **WordSmith Tools** or any of the individual Tools again if it's already running. Just

Alt-tab back to the instance which is running.

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.

XX days left

Message for limited period users only. At the end of this time **WordSmith** will revert to demo mode.

Index