MP 2.2
(Manual Update)

Michael Barlow

ATHELSTAN

Requirements:
Windows 95 or higher.
Disk space: **1 MB**
RAM: **32MB**

Related products:
ParaConc
Corpus of Spoken Professional American English
MonoConc 1.5
MonoConc Pro 2.0
Concordances in the Classroom. 1997. C. Tribble and G. Jones
Learning with Corpora. 2001. Guy Aston (ed.).

info@athel.com
www.athel.com
Tel: **(713) 523-2837**
Fax: **(713) 523-6543**

Athelstan
2476 Bolsover, Suite 464
Houston, TX 77005
U.S.A.

Preface

The descriptions in the following pages, taken in conjunction with the *MonoConc Pro* manual, provide the necessary information to work with *MP 2.2*. The assumption in writing this addendum is that you have read the *MonoConc Pro* manual thoroughly and that all want to hear about is what is new in *MP 2.2*.

If, given this description, you are feeling a little nervous, you might want to know that you can also check the online help available from the INFO menu.

Naming

As noted in the previous manual, *MonoConc Pro* is not the best name in the world. In what follows, we will follow the example of companies such as British Petroleum (BP) and Imperial Chemical Industries (ICI) and seek cover in the form of initials; hence, we will use the name *MP 2.2*.

New Features

It is relatively easy to add new features to a software program, but the continual addition of feature can obscure the flow and design of good software, bogging the user down in a swamp of options. The aim in creating *MP 2.2* is to add a couple of major components (Highlight Collocates and Corpus Comparison); to make adjustments to overcome an English-bias; and to make minor modifications to make the software more useful (such as allowing the concordance results to be saved to an html file), while maintaining the integrity of the original program.

Book

For those who would like a fuller description of text analysis, see *Corpus Analysis and Concordancing Using MP 2.2*.

Contents

1. Installation and Getting Started

1.1 Installation

copy files

MP 2.2 can be installed by copying all the files on the floppy disk or all the downloaded files to the hard disk on your computer. In most cases, the first step is to make a folder called MP and then simply copy the files to that folder. The files constitute the complete program; it is not an upgrade utility.

It is a good idea to create four subfolders within your MP directory: **Texts** (for different corpora), **Workspaces** (for workspace files), **Results** (for saved concordance and frequency information) and **Loads** (for stop lists and complex searches). You might also want to create a subfolder **Web** for results saved in html format.

For the program to function, the only file that actually needs to be transferred to your hard drive is **MP2.2.exe**, which is around one megabyte in size. The help file, **MP2.2.hlp** and the help contents file **MP2.2.cnt** should also be copied to the hard disk.

1.2 Requirements

1.2.1 Disk space

As noted above, the software files constituting *MP 2.2* require around a megabyte of disk space on a hard disk drive.

1.2.2 Compatible versions of Windows

MP 2.2 is a 32-bit program and hence must be run under Windows 95 or higher. Any higher Windows versions, including Windows 2000/XP/NT, are acceptable platforms. *MP 2.2* is not optimised for any particular system and is designed to run under a wide range of hardware/software configurations.

1.2.3 RAM

It is recommended that a computer with Windows 95 installed have 16 MB of RAM and a Windows XP system should have a minimum of 32 MB of RAM. As always, the more RAM the better, but the program should run under a minimal set-up.

1.3 Choosing a Language

Russian, Thai

The appropriate method of entering characters with accents and the definition of alphabetical (sorting) order depends on whether the current language is, for instance, English, Russian or Thai. Choosing LANGUAGE displays a list of languages, allowing the selection of the one appropriate for your corpus. (The range of languages displayed will depend on the version and configuration of Windows installed on your machine.) If the language you want is not present in the list, you should simply select the font that you want to use. Selecting the appropriate language/encoding is particularly important if you are working with CJK languages. See section 7.3.

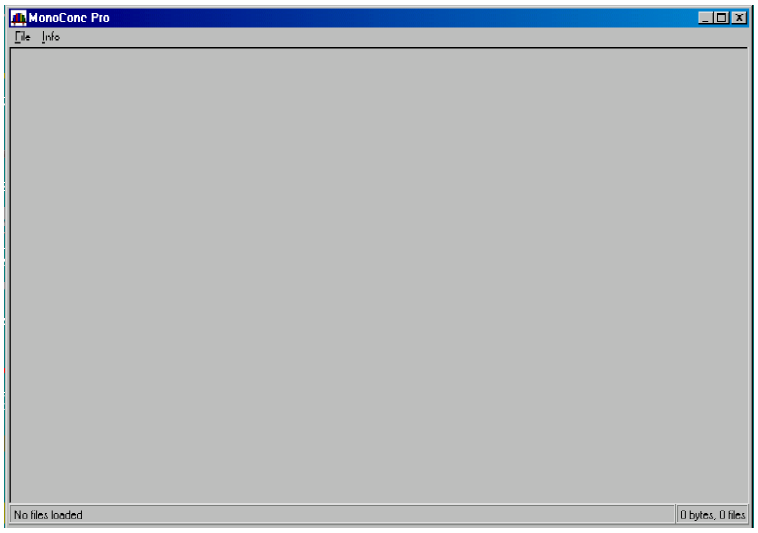
For information on entering accented and other special characters, see section 7.2.

Chinese

If *MP 2.2* is running under a system as different as, say, Chinese Windows, then the settings and behaviour will naturally be somewhat different from the description given here.

1.4 Getting started

The initial screen is shown below; the only difference from the 2.0 version is that the clock in the lower right of the window has gone, replaced by information on the size of the corpus and the number of files.



1.4.1 Count words

In the File menu is the option COUNT WORDS WHILE LOADING If this is checked, then when COLLECT TAG INFO is run and the corpus files have been processed the number of word tokens and number of word types will be displayed in the lower right of the window. This slows the loading process a little and so if a word count is not needed (or you plan to create a full word frequency list), then you can leave this command unchecked.

2. Corpus Frequency Lists

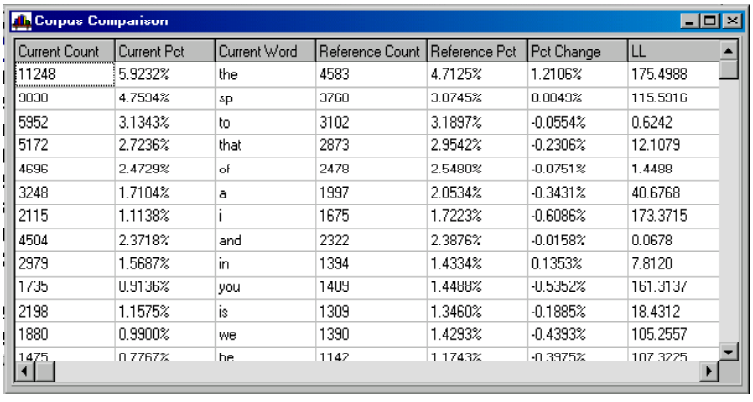
2.1 Corpus frequency list

The basic operation of creating a frequency list for a corpus is unchanged. The only difference is that the number of lines that can be saved (which equals the number of word types) is now increased substantially to 950,000.

2.2 Corpus comparison

Creating and saving a frequency list to file for a reference corpus allows the possibility of comparison with other corpora. Once the frequency list for the reference corpus has been saved to a file, create a frequency list for the current corpus. The CORPUS COMPARISON command then becomes available and allows the user to select the file containing the reference frequency list. (The frequency window must be active otherwise the CORPUS COMPARISON command will be greyed out.)

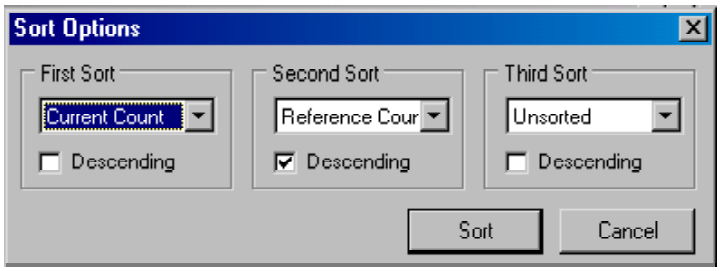
The program then compares the two frequency lists and produces a corpus comparison table, as shown below.



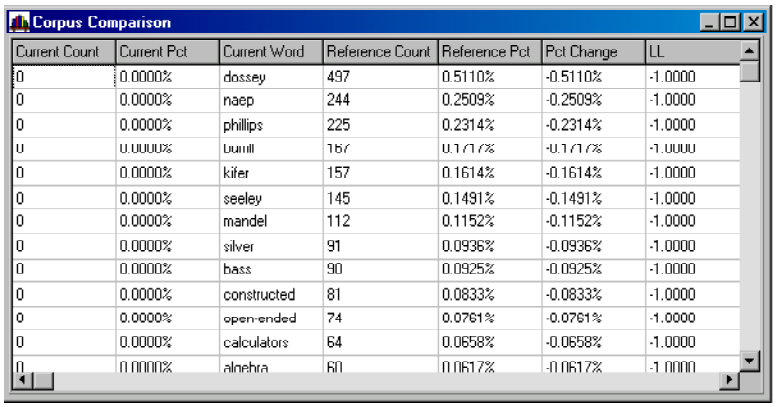
Current Count	Current Pct	Current Word	Reference Count	Reference Pct	Pct Change	LL
11248	5.9232%	the	4583	4.7125%	1.2106%	175.4988
9030	4.7594%	sp	3760	3.8745%	0.0849%	115.5916
5952	3.1343%	to	3102	3.1897%	-0.0554%	0.6242
5172	2.7236%	that	2873	2.9542%	-0.2306%	12.1079
4696	2.4729%	of	2478	2.5480%	-0.0751%	1.4488
3248	1.7104%	a	1997	2.0534%	-0.3431%	40.6768
2115	1.1138%	i	1675	1.7223%	-0.6086%	173.3715
4504	2.3718%	and	2322	2.3876%	-0.0158%	0.0678
2979	1.5687%	in	1394	1.4334%	0.1353%	7.8120
1735	0.9136%	you	1409	1.4488%	-0.5352%	161.3137
2198	1.1575%	is	1309	1.3460%	-0.1885%	18.4312
1880	0.9900%	we	1390	1.4293%	-0.4393%	105.2557
1475	0.7767%	he	1142	1.1743%	-0.3975%	107.3225

The columns in the table are Current Count, Current Percentage, Current Word, Reference Count, Reference Percentage, Percentage Change, Log Likelihood[1].

The table can be sorted according to one or more of these parameters. For example, to see the words occurring in the reference corpus, but not the current corpus, we can rearrange the table to show the words occurring 0 times. Choose SORT from the FREQUENCY menu with the settings shown below.



The second sort in this dialogue box will order the words in the reference corpus that were not present in the current corpus with the most frequent words at the top. The transformed table is shown below.



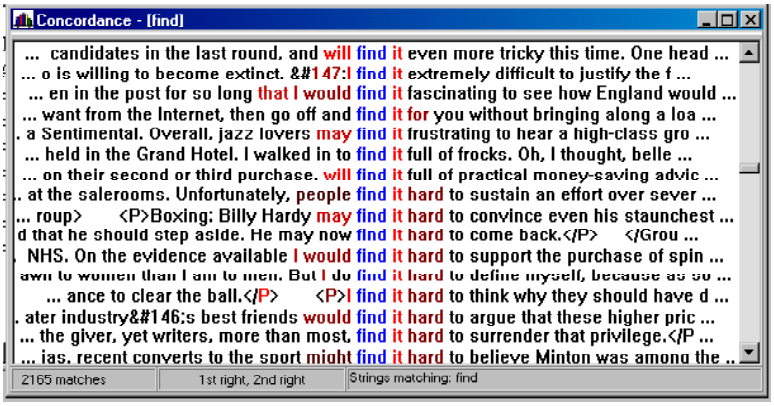
Current Count	Current Pct	Current Word	Reference Count	Reference Pct	Pct Change	LL
0	0.0000%	dassey	497	0.5110%	-0.5110%	-1.0000
0	0.0000%	naep	244	0.2509%	-0.2509%	-1.0000
0	0.0000%	phillips	225	0.2314%	-0.2314%	-1.0000
0	0.0000%	burrill	167	0.1717%	-0.1717%	-1.0000
0	0.0000%	kifer	157	0.1614%	-0.1614%	-1.0000
0	0.0000%	seeley	145	0.1491%	-0.1491%	-1.0000
0	0.0000%	mandel	112	0.1152%	-0.1152%	-1.0000
0	0.0000%	silver	91	0.0936%	-0.0936%	-1.0000
0	0.0000%	bass	90	0.0925%	-0.0925%	-1.0000
0	0.0000%	constructed	81	0.0833%	-0.0833%	-1.0000
0	0.0000%	open-ended	74	0.0761%	-0.0761%	-1.0000
0	0.0000%	calculators	64	0.0658%	-0.0658%	-1.0000
0	0.0000%	algebra	60	0.0617%	-0.0617%	-1.0000

[1] See P. Rayson and R. Garside. 2000. Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing Corpora. ACL 2000. Hong Kong.

3. Searching the Corpus

3.1 Searching

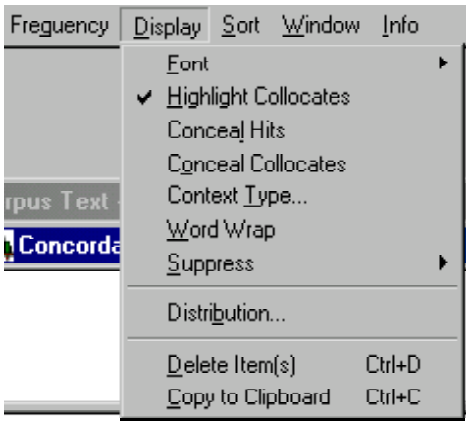
The search commands are unchanged but the results may look different, as illustrated in the screen shot below. (The differences may be hard to detect in the printed version of this update.)



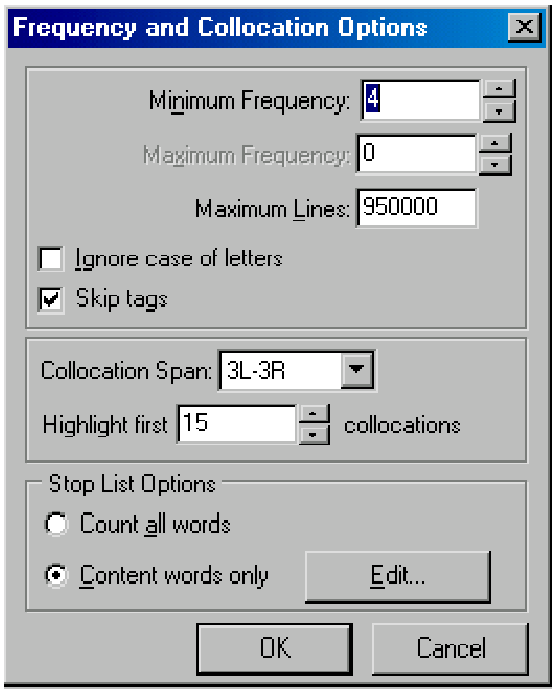
As I hope you can see, the frequent collocates of *find* are highlighted in colour. The frequent collocate *it* (immediately following *find*) is in bright red and *hard* (2nd right) is a duller red.

The highlighting can be turned on or off for any particular set of concordance results using the HIGHLIGHT COLLOCATES command in the DISPLAY menu.

In addition, there is a CONCEAL COLLOCATES command in the DISPLAY menu.

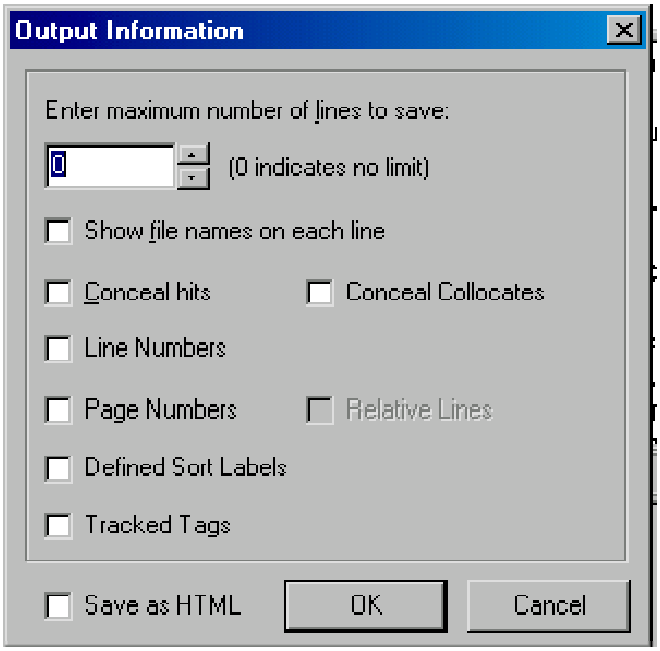


The settings controlling the display of the collocates are in FREQUENCY OPTIONS. The appropriate span of collocates, from 1L-1R to 4L-4R can be selected along with the number of collocates to be highlighted: the top 15, for instance, for each position (1L, 1R, etc.). If function words are to be excluded, select CONTENT WORDS ONLY and enter the stop words.



3.2 Saving the concordance lines

To save the contents of the concordance window, select SAVE AS FILE from the CONCORDANCE menu. A dialogue box appears and the name of the results file can be entered.



html — The default option is to save just the concordance results in the file. However, there are a variety of options concerning other information that can be saved along with the concordance lines, including SHOW FILE NAMES ON EACH LINE, CONCEAL HITS, LINE NUMBERS, PAGE NUMBERS, DEFINED SORT LABELS ,TRACKED TAGS, and CONCEAL COLLOCATES. The results may also be saved as an html file, which can then be displayed via the web. One advantage of doing this is that the colour information of the highlighted collocates is retained.

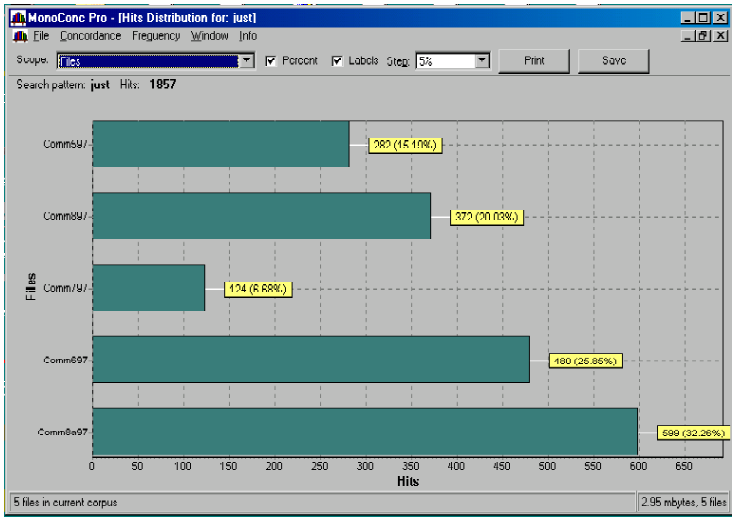
line-up — The alignment of the keyword in saved files has been improved in *MP 2.2*.

4. Distribution

skewed data

It is often useful to obtain a sense of the distribution of hits through a corpus.

The graph below shows the distribution of *just* in five files in the Meetings sub-corpus. The x axis indicates the number of hits and the y axis shows the individual files name. For other options, see the *MonoConc Pro* manual.

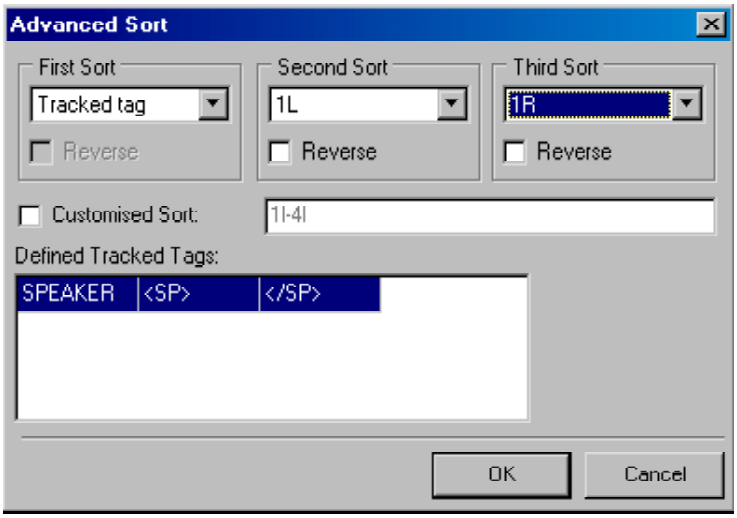


5. Sorting

5.1 Advanced sort

An ADVANCED SORT option is divided into two parts. The first, upper component of the dialogue box invoked by the selection of ADVANCED SORT allows the user to choose a primary, secondary and tertiary sort order (labelled as FIRST SORT, SECOND SORT and THIRD SORT).

A new parameter that can be used to sort the concordance lines is Tracked tags.



In the example shown above, the First Sort is defined in terms of the value of the Speaker Tag, which means that the concordance lines will be reordered in a way that clusters the hits associated with each speaker.

In this example, there is only a single tracked tag. If there are multiple Tracked tags, the one to be used in sorting can be selected from the DEFINED TRACKED TAGS box.

6. Collocates and Collocations

6.1 Collocate frequencies

Choosing COLLOCATE FREQUENCY from the FREQUENCY menu (or CTRL-F) displays the collocates of the search term ranked in terms of frequency.

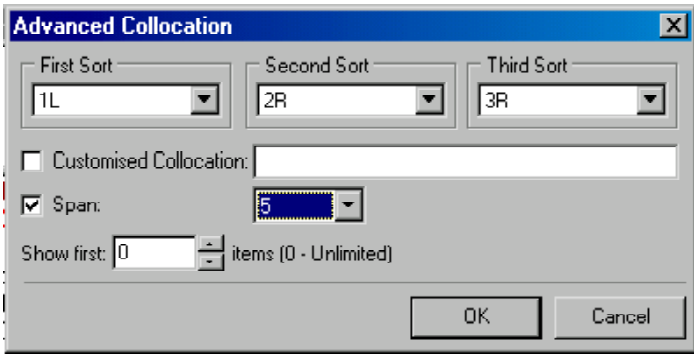
The span of the collocates displayed is set in FREQUENCY OPTIONS. The values are 1L-1R to 4L-R.

6.2 Advanced collocations

collocations

To calculate the frequency of collocations rather than collocates, it is necessary to choose ADVANCED COLLOCATION from the FREQUENCY menu.

A new component of ADVANCED COLLOCATION is SPAN, as illustrated in the dialogue box below.



The value of span can be set to between two and five words. With the above setting of 5, the program will generate a frequency list of 5-word clusters containing the search word.

7. Working with languages other than English

7.1 File format

The appropriate format for MP 2.2 files is ANSI (Windows text). If your non-English files are in ASCII format and you load them into MP 2.2, then the accents will appear garbled. You need to load the files into a Windows word processor such as Word, making use of the appropriate code page, and then save the files (with a different name) as text files.

7.2 Entering accented characters

Choosing the appropriate language before loading a corpus helps to make the software work appropriately. Choosing a language will not change the keyboard, however, making it difficult to enter special characters such as *ç* and *é* to give examples from French. In order to make it easy to enter these characters in the search string, a French virtual keyboard must be installed. To do this, select Control Panel under Settings in the Windows Start menu. Click on Keyboard and follow the directions for installing keyboards for different languages. Once installed, you can use the ALT to switch from one keyboard to the next.

7.3 Working with Chinese, Japanese, Korean

Chinese, Japanese, Korean, and other two-byte languages present particular problems for a concordancer, which is a word-based searching program. It is, nevertheless, possible to search CYK texts.

It is important to choose the appropriate language/encoding system before loading the files.

Segmented and non-segmented versions

In most representations, CJK are displayed without spaces between words and a special segmentation program is needed

to add spaces or other indicators of a word boundary in the appropriate places—a process that is not trivial. If the text has been segmented, then the ordinary text search will work, as long as the word boundary indicator is listed in the Word Delimiter textbox.

If the text is not segmented, then the simple text search option cannot be used and searching must be done using the regex search option, which is a part of the ADVANCED SEARCH command. This allows all the instances of the search string in a text to be displayed in a KWIC format, in the usual way.

That is the good news. The bad news is that the absence of word boundaries means that it is impossible to look at the collocates of the search string. The only manipulation of the search results that is possible is a right sort. A left sort will be meaningless in this situation since what is interpreted as a word boundary could be anywhere in the preceding string, although one option is to use the reverse option for a Left Sort in the Advanced Sort command. This will sort the characters based on the character immediately preceding the search string.

8 Tricks and Tips: Some final comments

As you get used to the different options and settings in MP 2.2, you will become more skilled in exploiting different working techniques to manipulate your results in the way that gets as close as possible to the desired goal. In this final section, we provide a miscellany of techniques for working with *MP 2.2.*

8.1 Counting the search term

If you perform a search containing some kind of wild card character, then you are like to retrieve different keywords or phrases. Thus a search for ***self** will find *self* plus *myself, yourself,* etc. How do we then get the program to count the individual instances of the words found? The answer is to use the ADVANCED COLLOCATION command. Choose this command and then set the First Sort to Search Item and the Second and Third Sort Items to None. Click on OK and the count of the different forms of *self* will appear in a results textbox.

8.2 Loading the search results as a corpus

There is no save-search feature that allows the retrieval of a concordance results window. The concordance results can only be saved as a file. You might want to recreate a particular concordance window so that, for instance, collocate frequency information can be extracted, and you can do this by loading the saved concordance file as corpus and searching for the search term again.

It may be necessary to adjust the context in order for the concordance lines to display properly and the context window and other links to the original corpus will not be available.

The search term is enclosed in square brackets when it is saved to a text file and it is safest to include the square brackets in the search query (which may involve taking the brackets out of the list of word delimiters in SEARCH OPTIONS).

If you search for the word without the brackets you have to be aware of the possibility of an overcount based on the following scenario. You search initially for a common word such as *thing*. Sometimes *thing* occurs twice in the same sentence, which causes no problem; each instance is displayed on a separate concordance line. If the results are saved with the sentence as the chosen context type, then we have a situation in which the same sentence is included twice, once for each hit. Again this is no problem, but if the saved concordance lines are loaded as a corpus, then there are two extra instances of *thing*. You can delete one of the pair of sentences or you can search for **[[thing]]** and not **thing**. Make sure that you have the same number of hits as you had in the original search.

8.3 Finding hapax legomena

To find all the words that occur only once in a corpus, you simply set both the MINIMUM FREQUENCY and the MAXIMUM FREQUENCY to 1 in FREQUENCY OPTIONS and create a frequency list, which will be displayed in alphabetical order. As always, you should check the word delimiters and the case-sensitive settings to make sure that the results are as required.

8.4 Creating a complete concordance

concordance

One meaning of a concordance is a complete listing of all the words in a text along with their context of use. To achieve this, it would be necessary to find all the words in a corpus and perform a concordance search for each one. *MP 2.2* is a query-driven concordance program and is not designed to produce a complete concordance of a text, but you might approach such a task in the following way. First create an appropriate corpus frequency list; save the list as a file (without the frequency information); modify the list as appropriate (by deleting unwanted words); and then load the file containing the list of words into BATCH SEARCH using IMPORT PATTERNS or using LOAD in the EDIT SEARCH PATTERNS dialogue box.

Note: if you have produced a corpus frequency list, it is best to quit *MP 2.2* before attempting another heavy processing task like this.

8.5 Extracting collocations

Collocations are tied to search terms and so there is no way to simply extract the collocations in a text. However, if you enter a group of words, such as an academic word list and have the program search for all the terms, you can thenuse the SPAN option in ADVANCED COLLOCATION to present all the two word (or three word, etc.) collocations in frequency order.

INDEX