Journal of the Chinese Language Teachers Association, Volume 37:2. May 2002.
Pages 1-58. (A pre-final version appears in the Journal.)

Concordancers and Concordances: Tools for Chinese Language Teaching and Research

Marjorie K.M. Chan
The Ohio State University

This paper presents an introduction to concordancers, and to the concordancing of Chinese e-texts in particular. Demonstrations are given of searches using spaced and non-spaced source e-texts, with the concordance results presented in Keyword-in-Context (KWIC) display format. There are illustrations to accompany discussions of full-text concordances, and of concordances targeting specific words or phrases. The writer suggests how concordancers might be used in language-teaching and in conducting research on various linguistic phenomena of the Chinese language. An appendix compares several concordancing programs capable of handling Chinese e-texts.

1. Introduction[1]

Concordancers are tools developed for mainframes, PC's, Mac's, and other computers and operating systems for conducting searches for words, or strings within a word, and then, in a matter of seconds, exhaustively listing the occurrences of that word (or string) in the electronic corpus, together with the contexts in which the words or strings occur in the source text. Concordancing software, or concordancers, thus provide an easy and yet powerful means to study the multiple meanings and functions of a given word, compare usages and distribution of two or more words that are near-synonyms, analyze vocabulary choices and grammatical patterns in different portions of a literary work to determine single or multiple authorship, generate collocations for studying words commonly associated with the searched word, and so forth.

Concordancing software programs have been used in foreign language classrooms in the past decade for English and other Western languages, but the same is not true for learning Chinese. On the World Wide Web, there are ample electronic texts (e-texts) encoded in Big5 (with 'Big5' used here to refer to Big5 and Big5+) and GB (with 'GB' used here to include both GB and GBK) that can

[1] A preliminary introduction to 'Concordancers, Concordances, and Chinese Language Teaching' was presented at the 2001 Annual Meeting of the Chinese Language Teachers Association in Washington, D.C. My thanks for the questions and comments received afterwards.

serve as Chinese-language corpora. Such e-texts include short stories and novels, e-news items, e-texts prepared for language learners, as well as numerous other GB/Big5-encoded electronic texts. However, despite the explosion of available digital materials, there remains a lag in using concordancing programs in the Chinese field. This is true whether one is referring to literary studies of intertextual connections, corpus-based linguistic research on distribution of lexical, grammatical, and discoursal phenomena, or language-textbook and pedagogical-materials preparation. There are a number of reasons for the lag in using concordancers in Chinese; the most important pertain to technical problems in dealing with double-byte Chinese characters and non-spaced Chinese texts, which make such tasks as searching, wordlist-generation, creation of concordances (in key-word-in-context (KWIC) format), and sorting a major hurdle. This is particularly true if one wishes to use concordancing programs that were designed for English orthography originally, and then extended for other, left-to-right scripts that use single-byte encoding. The situation is little different from the lag in using computer technology for Chinese linguistic research in general; whereas computers were used as early as the 1960's for linguistic analyses of English (Kennedy 1998:93), for most Chinese linguists, it was probably not until the early 1980's, when micro- and mini-computers became popular, that desktop computers, with software for handling Chinese characters, were harnessed for linguistic research.

Corpus-based Chinese linguistic research, such as Huang and Ahrens (1999) on post-verbal *gei* 給 and Tao (2000) on adverbs, make use of electronic corpora and searching/concordancing software, while others might use something yet simpler, such as a DOS-based word-counting program, such as Wang (1999) in the study of adverbial clauses in recorded conversational transcripts. In the case of corpora that are not in digital format, manual checking and counting is needed, such as the study of Cantonese sentence-final particles in Chan (1996, 1998), where every occurrence of *je* and *jek* was manually searched in the hardcopy transcripts and tabulated. Whereas human beings are slow and can make errors in counting, computers can perform these searching tasks swiftly *and* accurately.

Concordancing per se existed long before computers were used for concordancing of texts. In Europe, concordancing efforts were underway since the Middle Ages, especially for biblical studies, and then extended to classical texts such as the works of Shakespeare and other well-known literary giants. Perhaps the earliest known concordance was that completed in 1230 for the Latin Bible, prepared by Hugo de Saint-Cher (or Hugo de San Charo), with the aid of 500 fel-

low-Dominicans.[2] And the use of corpora for lexicography goes back at least to the early seventeenth century. One monumental work that was compiled without the aid of computers was the first edition of the *Oxford English Dictionary*, with the first volume published in 1884 and the twelfth and final volume in 1928. Work on that dictionary spanned 71 years, with some 2,000 volunteer readers collecting about five million citations (totalling 50 million words) to illustrate meaning and usages in the 414,825 entries in that dictionary. The third edition of *New International Dictionary*, published in 1961, may have been the last major English dictionary to be completed without an electronic database (Kennedy 1998:14-15).

This paper is organized into seven sections; after this initial section, section 2 briefly introduces the use of concordancing for English e-texts before proceeding to the main focus of this paper, namely, section 3 on concordancers and concordancing with Chinese e-texts, including the use of non-spaced e-texts versus character- and word-spaced e-texts, and plain e-texts versus those with parts-of-speech tags. Issues concerning encoding of Chinese e-texts are also addressed. The remainder of the paper follows, with section 4 on concordancing and Chinese language teaching, section 5 on sources and preparation of Chinese e-texts, section 6 on concordancing programs for Chinese e-texts (with comparisons of four concordancers placed in the appendix), and section 7 concluding the body of the paper.

2. Concordancers and Concordancing of English E-Texts

A simple concordance of an English e-text as corpus is illustrated in Figure 1, displaying the results of a search for 'handsome' in Jane Austen's 1815 novel, *Emma,*[3] using the Windows program, *MonoConc Pro 2.0.* The word 'handsome' has multiple meanings and usages;[4] a handsome letter or reply refers to appropriate speech or wording, and can imply gracefulness of style. It is used in reference to a generous gift from a certain Colonel Campbell, whose income (by pay

[2] See Tribble, Chris, and Chris Jones (1990) and the University of Glasgow's Humanities Advanced Technology and Information Institute's web page, 'Lecture 8 - Concordances, Thesauri, Indexes' at <www.stu.hatii.arts.gla.ac.uk/Courses/textproc/restricted/TPLect8bConcord.htm> (accessed on 19 December 2001). See also 'Catholic Encyclopedia: Concordances of the Bible' online at: <www. newadvent.org/cathen/04195a.htm> (accessed on 19 December 2001).

[3] The e-text was downloaded from Project Gutenberg's FTP site at: <ftp://ibiblio.org/pub/docs/books/gutenberg/etext94/emma10.txt>. The website for Project Gutenberg, with search engine for locating e-texts archived at the site, is <promo.net/pg/> (accessed on 19 December 2001).

[4] *Oxford English Dictionary Online* <dictionary.oed.com/entrance.dtl> contains six main entries (with several sub-entries) for 'handsome' as an adjective.

and appointments) is also rather handsome; that is, the income is of a considerable sum. One or two rooms in a house (the Abbey) are described as handsome—that is, conveying the sense of rooms of a fairly large size and presumably rather grand and formal—contrasting with the rest of the (many) rooms in that house that are merely 'comfortable.' And in reference to a person, one who is handsome is one with a fine form or figure (and usually with respect to full size or stateliness); thus, in Jane Austen's time and into Victorian England, both men and women may be described as handsome.

Figure 1. Search for 'handsome,' in Jane Austen's *Emma* (1815).

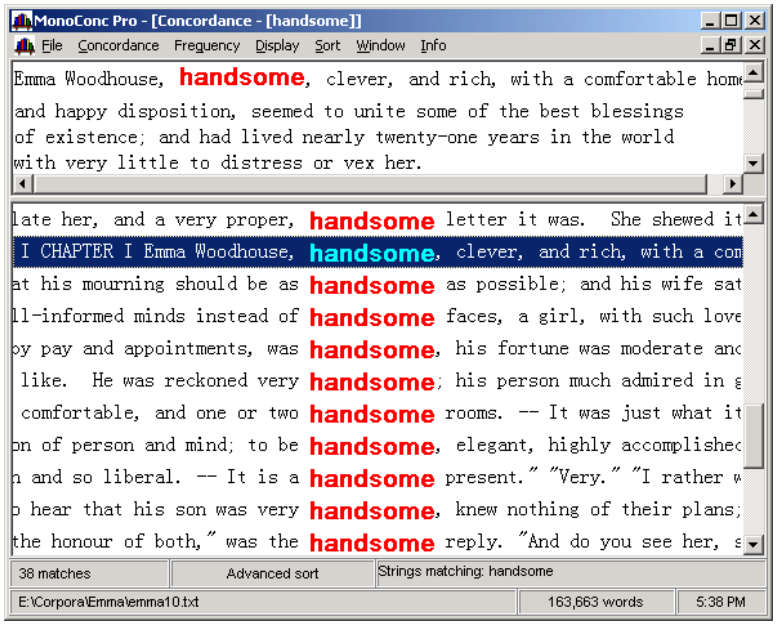
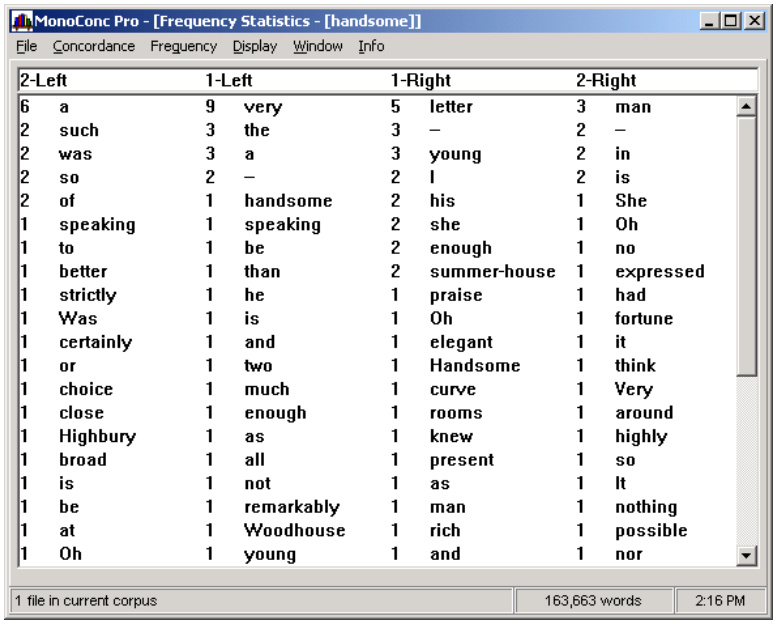


Figure 1 contains two vertically-arranged windows. The lower window contains a KWIC (Keyword-in-Context) display of the search results for one keyword, namely, 'handsome.' As shown at the bottom-left corner of the program, the corpus contains 38 occurrences ('matches,' 'hits,' or 'tokens', as in 'type-token' distinction) of the keyword. Eleven of the occurrences are displayed in the lower window of the figure, with each token of the keyword highlighted, and the context to its left and right displayed as well. At the same time, the entire line of the second token of 'handsome' is also highlighted. The full context of that token

of the keyword is given in the upper window, highlighted to show exactly where in the source e-text that token is located.

Alternatively, one can take a quick look at possible collocates of 'handsome' in the corpus, as shown in Figure 2, where the left and right collocates of 'handsome' are displayed, sorted by frequency. Among the right collocates that are nouns, one sees 'letter' (occurring five times), 'summer-house' (occurring twice), 'rooms' (occurring once), and so forth. Hence, this alternate view provides another convenient approach for studying the searched word.

Figure 2. Collocates of 'handsome' in Jane Austen's *Emma* (1815).

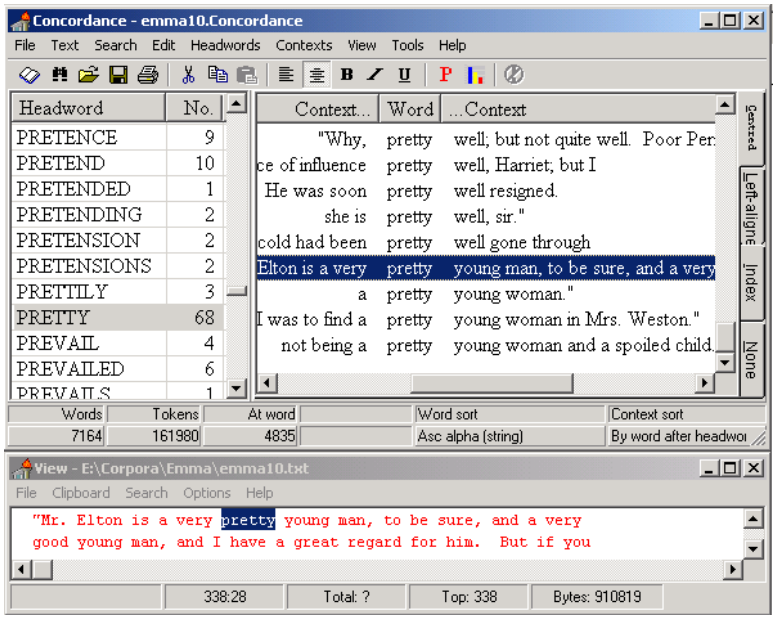


Lines in concordances generated by concordancing software can be deleted if they are not relevant for the search (e.g., deleting all occurrences of 'handsome' that were used in *Emma* to describe human beings). Concordances also have powerful sorting capabilities, so that contexts to the left and to right of the keyword can be sorted to display frequently-occurring collocates and distributional patterns that might not be obvious by scanning an e-text with the human eye, without the aid of a concordancer. For example, in *Emma,* while both 'girls' and 'women' may be described as 'pretty,' only women are described as 'arrogant,'

'conceited,' 'respectable,' 'fine-looking,' 'elegant,' and 'charming,' and only girls are described as 'insolent,' 'headstrong,' 'foolish,' 'sweet,' 'good,' and so forth. The adjectives, 'great,' 'fine,' and 'noble,' are reserved for 'ladies' in this novel.

The more powerful concordancers can also handle multiple keyword searches and full concordances; that is, a concordances of *every* word in the corpus. Figure 3 displays an example of a full concordance of Austen's three-volume novel, *Emma*, containing .16 million words.

Figure 3. A full concordance of Jane Austen's *Emma* (1815).



The concordance of the e-text was completed in 11.47 seconds using *Concordance 3.0*,[5] yielding over 7,000 unique words that are all alphabetically

[5] The full concordance of this three-volume novel, together with sorting of all the keywords, was accomplished in 11.47 seconds using my 500 MHz, 589 MB of RAM, Pentium III notebook, and with the concordancing program running under English Windows 2000. Concordancing speed is dependent upon a number of factors: these include size of the corpus, including number of unique words and tokens (i.e., total word count); computer speed; amount of RAM; amount of context to display (e.g., number of words before and after the keyword); and the program's speed in executing the commands for retrieving, sorting, and displaying of the results. Unless noted otherwise, the remaining figures in

sorted. The speed and efficiency of using concordancing software are obvious: manually concordancing all the words in the novel would have been a drudging task that could take days, not seconds, minutes, or even hours. Multiply that by a corpus consisting of, say, all six of Jane Austen's novels. Tasks that would be too daunting if done manually can be contemplated and carried out using concordancers. With improved computer technology and increased retrieval, sorting, and displaying capabilities come new questions that can be posed, limited only by one's imagination. With concordancing software at this stage of computer and software development, a full concordance of the Austen's novel can be completed and displayed in seconds, a far cry from the 500 monks' labor-intensive work in making a full concordance of the Latin Bible in the thirteenth century.

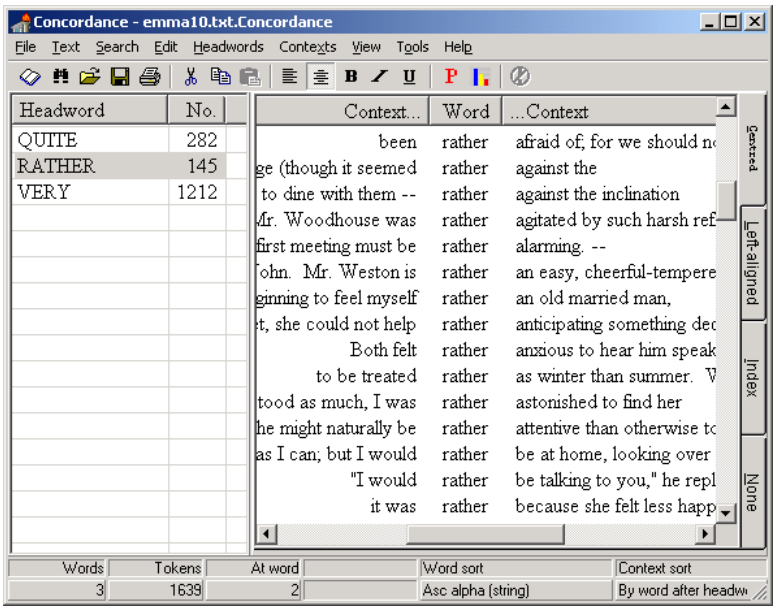
In Figure 3, three windows are shown, with the top two windows displayed automatically when the concordancing is completed. The Headword window on the upper left shows the full set of keywords (or 'headwords') in the e-corpus; that is, every keyword (with upper/lower-case ignored) appears in that window and all the keywords are sorted alphabetically. The highlighted keyword is 'pretty,' with 68 occurrences in the corpus. (The program can also optionally display percentages in a column next to the frequency figures column.) The upper-right window is the Context window, in KWIC display format, with the tokens of the keyword in the middle, flanked by the left and right contexts, with the amount of context to display selected by the user before making the concordance. (In this case, actual line was selected for the KWIC display, an option that is especially meaningful for concordances displaying lines of poetry.) The small, long window on the bottom, which can be moved separately and resized, is the View window, displaying a small portion of the source e-text where the specific token of 'pretty' is highlighted in the source e-text, containing a fuller context than the Context window can display. As shown in this figure, 'pretty' is not reserved for 'girls' and 'women'; it is also used with 'young men' (but not with grown, adult 'men,' nor with 'gentlemen' in the novel). Observe also that 'pretty' is used both as an adverb and as an adjective. Since the corpus is not tagged for parts of speech (i.e., a POS-tagged corpus), a search for 'pretty' indiscriminately retrieves all instances of 'pretty,' both adjectives and adverbs. To retain only the adverbial usage or only the adjectival usage, the unwanted tokens can be discarded by deleting them. The concordance results can be saved with the '.concordance' file extension in this software program, and re-opened later for further study.

---

this paper will be produced from screenshots of concordances using Rob Watt's *Concordance 3.0* program in English Windows 2000.

Multiple keyword searches ('fast concordances' in *Concordance 3.0*) are also possible, as shown in Figure 4, where only three keywords are searched for comparing their distribution, namely, the adverbs 'quite,' 'rather,' and 'very.' A right-sorted concordance is displayed in the Context window; that is, sorting is done on the word immediately after the keyword.

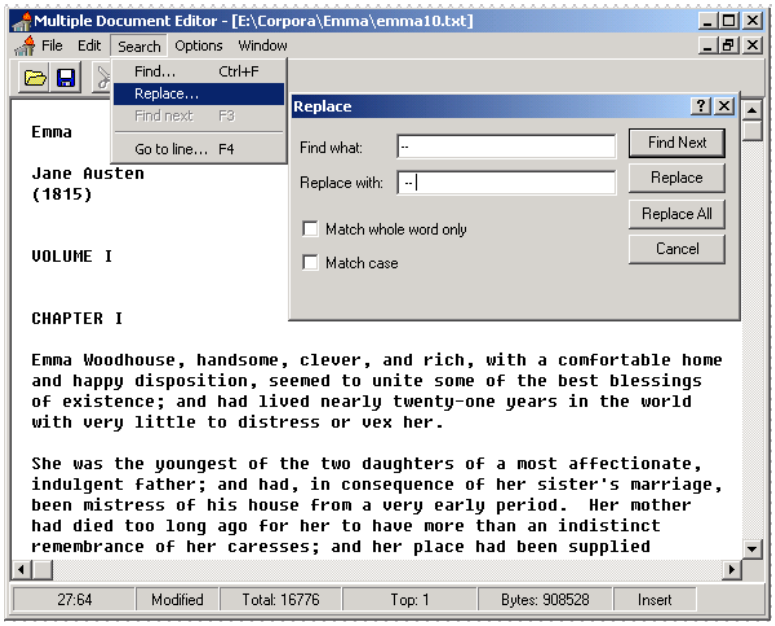
Figure 4. A fast concordance of 'quite,' 'rather,' and 'very' in Jane Austen's *Emma* (1815).



The source e-text is simply an ASCII file, as shown in Figure 5, where some introductory material on the creation of the e-text was deleted and some other minor preparatory work was done, such as finding two hyphens (--) and adding a space before and after them to ensure that these hyphens would not be treated as part of a word in the concordancing process. Punctuation marks are ignored as the default setting by the concordancers for English, so that a word ending in a comma or period, for example, would not be treated as a separate keyword from the same word occurring elsewhere in the text with a space after it. At the same time, punctuation marks are ignored and not included as separate keywords. As we shall see in section 3, retaining punctuation marks can be useful for obtaining

frequency data on sentence types (declaratives, interrogatives, and exclamatory sentences) in an e-text.

Figure 5. The e-text for Jane Austen's *Emma* (1815).



This section provides a glimpse of concordancing in English using concordancers developed for English e-corpora. Beyond the basics are complex sorting routines for morphological studies, marking-up of source e-texts for more refined searches, etc. Some of these capabilities of concordancing software will be introduced in section 3.

3. Concordancers and Concordancing with Chinese E-Texts

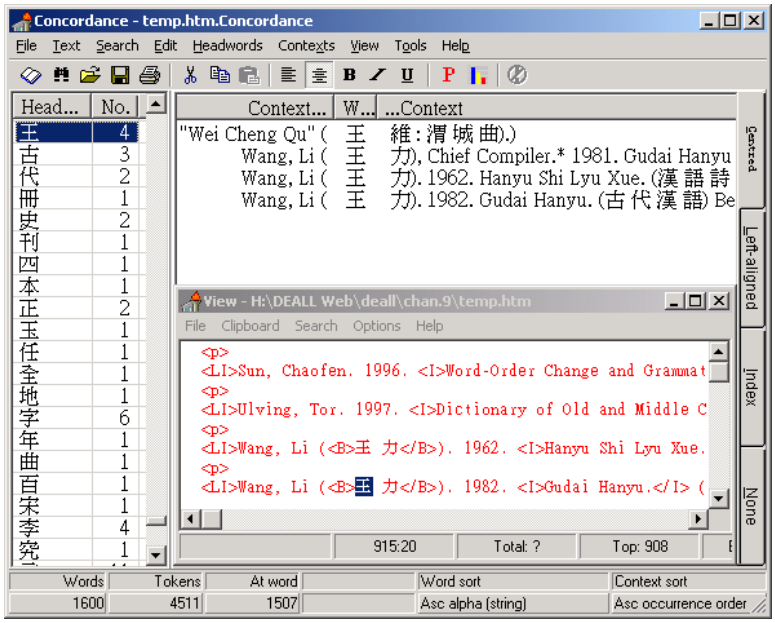
In this section, the types of source e-texts for Chinese concordancing are discussed. Section 3.1 briefly discusses encoding and e-texts in different formats, namely, plain text files, HTML-tagged e-texts, and e-texts that are formatted and word-processed (in RTF and MS Word DOC format), that are used as electronic corpora for concordancing. Sections 3.2 through 3.4 deal with different kinds of spacing (or non-spacing) of the e-texts. Section 3.5 discusses briefly e-corpora

that are tagged for parts of speech (POS-tagged) and used for corpus linguistic research, computational linguistics, and natural language processing.

3.1. Encoding and Chinese E-Texts for Concordancing

Plain e-texts are those Chinese-encoded texts with no formatting and no HTML-tagging, as found in web pages. While a program such as *Wenlin* (for PC's and Mac's) can handle a multitude of encoding systems for Chinese, including HZ, GBK, Big5+, Unicode, UTF8, and so forth, this is not typically the case with dedicated concordancing programs at this time. The usual encoding systems that general (i.e., Latin-script-based) concordancers can handle—if they can handle double-byte at all—are GB (and GBK) and Big5 (including extensions if supported by the font). Programs designed in recent years for concordancing work have built-in functions that enable the user to suppress HTML-tagging in making and displaying a concordance.

Figure 6. HTML-tagging suppressed in the concordanced results.



Thus, even though an author's name appears in bold in an HTML file, as shown in Figure 6, all HTML tags are suppressed in the KWIC display in the Context window. In Figure 6, the surname, 王 Wang, is the keyword highlighted as the top word in the Headword window, while in the Context window, the fourth token is highlighted. The keyword is also displayed and highlighted in the bottom-right, View window containing the source e-text, where one can see the tags, <B> and </B>, which flank the Chinese characters. As shown there, all the HTML tags have been retained in that source e-text, although they do not show up in the Headword window or in the Context window. The tags, with the opening marker, '<', the closing marker, '>', and all text contained between those markers, were suppressed at the time of creating the concordance. Note also that the Chinese characters in that web page had been input with spacing between characters; hence, each Chinese character was treated as a separate keyword in this full concordance, from a source web page that contains only a handful of Chinese characters.

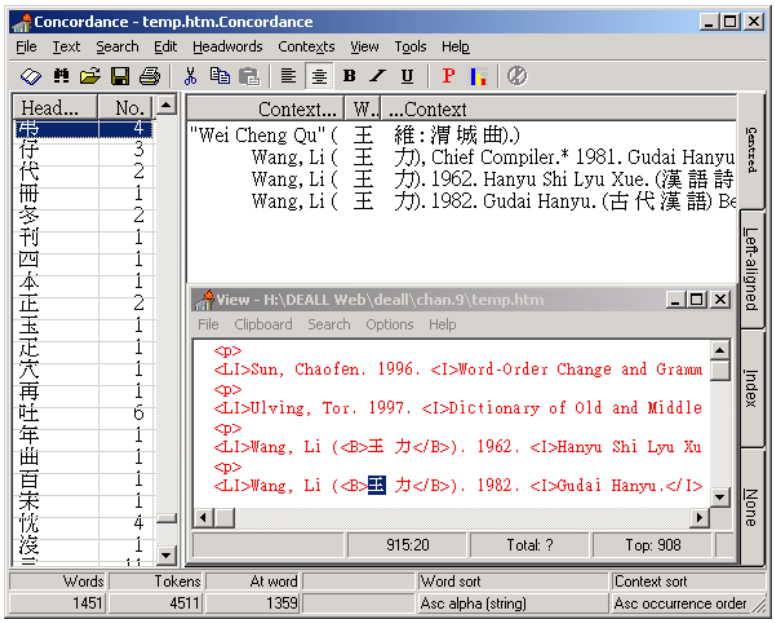
Instead of retaining the HTML tags, one could, alternatively, strip the files of all HTML-tagging. That would help to reduce bulk, especially in cases of large and heavily-formatted HTML files. Web-editing software typically can strip files of HTML tags in one quick and easy step. In addition, some concordancing programs, such as *Concordance 3.0*, and some Chinese software utilities, such as the 'Universal Code Converter' in *NJStar Communicator 2.x,* can also handle this task.

Concordancing programs typically cannot use e-corpora that are formatted by word-processing programs, although there are programs, such as *ConcApp 2.0,* that can read rich-text files (RTF files) and Microsoft word-processed (DOC) files as source e-corpora. Generally, for concordancing, files that are ASCII or Big5/GB-encoded plain texts, with or without HTML-formatting, are the most versatile for concordancing.

Note that even when the plain texts are used, whether a concordancer can handle texts with double-byte Chinese characters depends critically on whether the program has an option for selecting case-sensitivity, so that a word beginning with upper-case 'H' and one beginning with lower-case 'h', for example, would be treated as distinct keywords in a concordance. Without that option, the concordance would incorrectly display the encoding of Chinese characters, as each byte of a two-byte character needs to be treated as a separate, distinct entity; otherwise errors would occur, as shown in Figure 7. The same source file and the

same keyword as displayed in Figure 6 is used here, except that this time the concordance was made without taking 'case-sensitivity' into consideration. That is, the sorting order is identical to that in Figure 6 (i.e., the default, 'alphabetical,' sorting order), but the character that should be 王 Wang is incorrectly displayed in the Headword window as 弔 diào (alternate form of 吊 diào 'to hang'). It is in the Context window and the View window, the latter showing the same e-text as that in Figure 6, that 王 is correctly displayed. Other incorrect displays of keywords in the Headword window can also be observed.

Figure 7. Errors generated when case-sensitivity is not selected in creating a concordance.



Given the crucial need for handling what in English concordances would involve 'case-sensitivty,' a concordancing program that superficially appears totally incapable of handling double-byte CJK (Chinese/Japanese/Korean) characters may only need some 'tweaking' of the program to enable 'case-sensitivity' and, hence, proper display of the double-byte CJK characters. With the preceding background, we now proceed to discuss how non-spaced Chinese e-texts can serve as e-corpora for concordancing.

3.2. Using Non-Spaced Chinese E-Texts

Non-spaced, Chinese-encoded e-texts can be used in concordancing programs that are Latin-based, and designed for single-byte glyphs such as the letters of the English alphabet. However, one needs to search for strings of text via searches using regular expressions. That is, with regular expressions, the program merely tries to match the sequence of symbols that are being searched, without regard to whether that string of symbols is a 'word' or part of a 'word.' For English e-corpora, that allows for searching for a stem along with various derivations that include the stem plus any prefixes or suffixes, etc. A search for 'handsome' as a regular expression in Austen's *Emma* would have yielded 45 hits, including four occurrences of 'handsomely' and three occurrences of 'handsomest.' For Chinese-encoded text with no spacing, that same searching function can be used to search for Chinese characters as though they are simply a string of symbols. The *Wenlin* software program for learning Chinese has searching and concordancing capabilities that resemble regular expression searches; for example, as a search of the *Emma* e-text for 'handsome,' ignoring case, yielded 45 hits—37 cases of 'handsome,' one case of 'Handsome,' 4 cases of 'handsomely,' and three cases of 'handsomest.' That is, *Wenlin* can ignore case (abc=ABC) and, crucially, treats simple forms (partial words) and full forms (words) as matching.

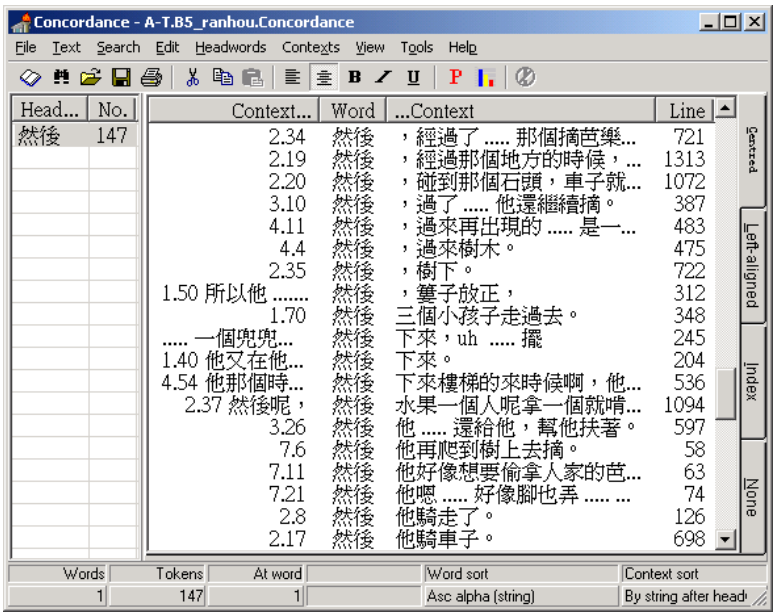
An example of using regular expressions to search for Chinese characters in non-spaced e-texts is given in Figure 8. A concordance was made of the conjunction, 然後 *ránhòu* 'and then, after that,' using the twenty transcripts (A through T) from Mary Erbaugh's set of Big5-encoded 'pear' stories. These transcripts contain oral narratives that recount the sequences of simple events in the *Pear Film*, produced under the direction of Wallace Chafe (then at the University of California at Berkeley).[6] The film is short—only seven-minutes—and has sound but no dialogue.

As shown in Figure 8, there were 147 instances of the keyword in the e-corpus. In the Context window, sorting was performed on the string to the right of the keyword. The corpus shows that 然後 is frequently used by the subjects to link sequences of events in their oral narration. By contrast, the conjunction, 结

[6] These are Big5-encoded transcripts of the 'pear' stories (Erbaugh 1990), from oral narratives produced by nineteen subjects (all females, ranging in age from eighteen to twenty-seven) in Taipei, Taiwan, in 1976, after watching the *Pear Film,* a film that has been extensively used for cross-linguistic studies of oral narrative production (see Chafe 1980). One additional transcript included in this e-corpus, Transcript T, was from a narrative produced by the interviewer for the study. Special thanks go to Mary Erbaugh for her generosity in making her transcripts and audiotapes available to me.

果 *jiéguǒ* ‘consequently,’ was used far less frequently: only 40 occurrences showed up in that e-corpus. This is at least in part because, unlike the use of 然後, the connection between two events linked by 结果 is necessarily tighter, since the second event is the result of the first. Furthermore, it may be that 然後 is not always used as a conjunction per se in these oral narratives; instead, at least some, if not many, of the tokens are in fact functioning as discourse markers or pause fillers, so that they occur far more frequently in a spontaneously-produced oral narrative than one would find in a written narrative of a story. Another conjunction, 後來 *hòulái* ‘afterwards, later,’ is used even less frequently than 结果; there are only 26 instances in that set of narratives. The choice and use of conjunctions in oral narratives could be studied more closely in a more thorough investigation of the discourse structure of these narratives.

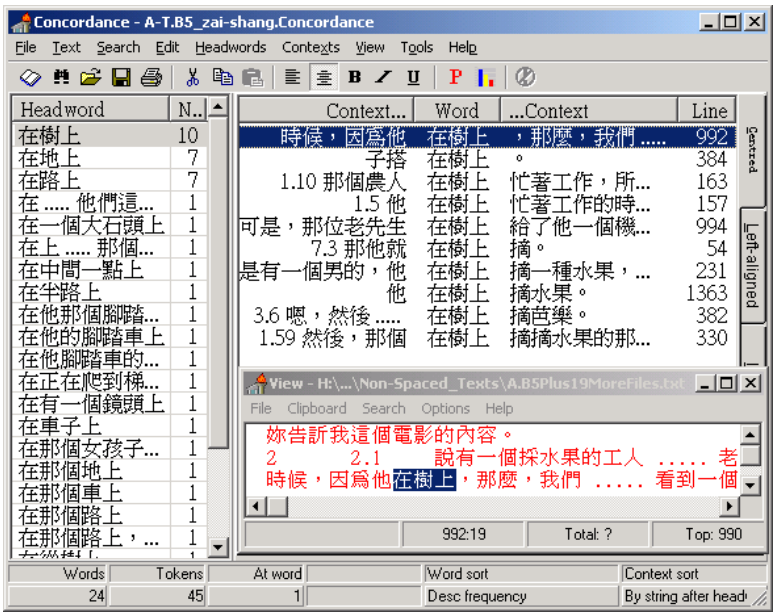
Figure 8. Search for 然後 in the Pear Stories (non-spaced texts).



With regular expressions, one can also conduct searches using ‘wildcards.’ However, note that concordancing programs differ significantly in their searching capabilities. Thus, only some concordancing programs, such as *Concordance 3.0,* can handle searching of non-spaced e-texts using wildcards. In Figure 9, a search was made using the twenty transcripts of the Pear Stories to search for all

phrases containing 在 *zài* 'at' and 上 *shang* 'up,' with a mono- or poly-syllabic noun (phrase) in between; that is, the pattern being studied is that of locative phrases containing '在 + NP + 上.' In conducting the search, an asterisk (*) was used as the wildcard for zero to any number of symbols between 在 and 上. (The exact search sequence in this program is the portion between the quotation marks: '在.*上', where a period followed by an asterisk is placed between the two Chinese characters.)

Figure 9. Search for 在 ... 上 using wildcards in the Pear Stories (non-spaced texts).



The search yielded 45 hits in total (shown in the bottom-left corner of the figure as '45 tokens'), of which there are 24 unique phrases (displayed as '24 words'). The phrases in the Headword window were sorted by frequency in descending order. (Other sorting routines in the program include sorting alphabetically, or sorting by length, word endings, or order of occurrence.) As shown in the Headword window, frequently-occurring phrases in these narratives are: 在樹上 *zài shùshang* 'on the tree' (10 hits), 在地上 *zài dìshang* 'on the ground' (7 hits), and 在路上 *zài lùshang* 'on the road' (7 hits). If the phrases were re-sorted by word-ending (using string sort), placed in the same sequence with 在路

上, for example, would be the following three phrases: 在半路上 *zài bànlù-shang* 'midway, on the way,' and 在那個路上 *zài nàge lùshang*. Concerning the last phrase, it might be noted that the subjects used the classifier 個 *ge* very frequently—in fact, a total of 861 hits in the corpus—including five times for 路 'road,' while the classifier 條 *tiáo* was used only four times in the entire corpus, and only really once for 'road.' One other case involves one of the subjects vacillating between 條 and 個, but ultimately choosing 個, as shown in the highlighted classifiers in the following excerpt, depicting the scene at the point where the small boy in the film had ridden his bicycle onto a small, dirt road (i.e., unpaved, country road): '走到了一**個條** . . . 一**個**小路上的 . . . 走到了一**個**路上的時候'.

Sorting 'alphabetically' may not seem meaningful when one is using Chinese e-texts; nonetheless, sorting of keywords and sorting of tokens of keywords, allow the same Chinese characters to be grouped together for observing distributional patterns and conducting quantitative studies.

The preceding examples show how non-spaced Chinese texts could be searched for words or phrases using a concordancing program that was initially developed for concordancing of English e-texts. In addition to such programs as *Concordance 3.0* and *MonoConc Pro 2.0,* which were developed for concordancing of single-byte, Latin-based e-texts but can also be used for concordancing of double-byte non-spaced e-texts, there are also some programs that were developed exclusively for handling double-byte Chinese, or at least include handling of double-byte Japanese as e-texts for concordancing. Two will be mentioned here: one is the *Wenlin Software for Learning Chinese,* or *Wenlin* for short (versions 2.0 and up include John DeFrancis' *ABC Dictionary*), for DOS, Windows and Mac operating systems. But because the *Wenlin* program is not a dedicated concordancing program, searching does not support wildcards, and keyword search results cannot be sorted, and are not given in KWIC display format with the keyword centered in the display. Nonetheless, for someone who already owns that program, concordancing is then one more useful function. At the present time, only *Wenlin* can search Unicode and UTF-encoded e-texts.

The second concordancing program is freeware that runs under Windows 95/98/NT and is designed for concordancing of Chinese or Japanese e-texts (in addition to English and French ones), namely, *ConcApp 2.0*. It does not have the full range of searching and sorting functions that commercial concordancing pro-

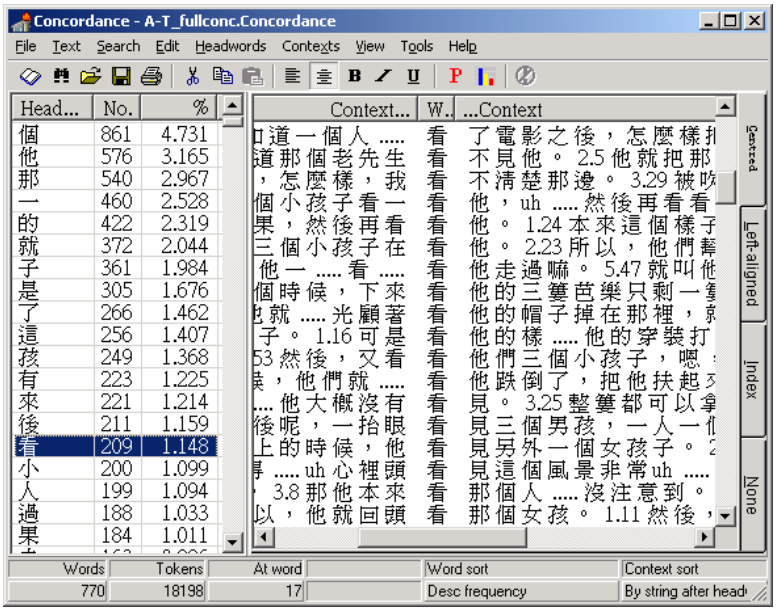
grams have, but it is a handy and useful tool for conducting simple searches. At the same time, since it 'recognizes' double-byte Chinese characters as words, statistics on the number of unique words in a given corpus of non-spaced e-texts can be obtained, together with a full list of the Chinese characters in the corpus and their frequency. For more specific information and a comparison of the four programs mentioned here (*Wenlin, ConcApp, MonoConc Pro,* and *Concordance*), see the appendix.

3.3. Using Character-Spaced Chinese E-Texts

Once we move to e-texts with spacing between Chinese characters, the opportunities for concordancing increase immensely. Programs that cannot handle searches using wildcards in non-spaced texts will be able to do so with character-spaced e-texts. Proximity searches and other searches involving 'word' count become possible. In addition, there are at least two other important advantages that can be gained in using character-spaced e-texts instead of non-spaced ones: one is the ability to display collocations (as in Figure 2 for 'handsome' in Austen's novel, *Emma*); two is the ability to generate full concordances (as in Figure 3 for Austen's novel).

Let us begin with the second case from which to discuss collocations. Spacing of e-texts opens up the possibility of producing full concordances, as will be illustrated here using the twenty non-spaced e-texts of the Pear Stories introduced in section 3.2. The source e-texts must first be formatted using a utility called a segmenter which "segments" or separates any two otherwise contiguous Chinese characters by inserting a space (ASCII number: 32) after each such character; one such utility is *NJStar Communicato*r's Universal Code Converter, which was used here. The new set of e-texts then served as the input to the concordancing program for making a full concordance, which took only 8.59 seconds to complete in *Concordance 3.0* (see footnote 5 on computer hardware). The results are shown in Figure 10, with keywords sorted by frequency. The Chinese character 個 *ge* has the highest frequency in this corpus, and 個 *ge* is also the most frequently-occurring classifier in these transcripts. Observe that 個 is followed in frequency by other Chinese characters that students encounter immediately or very soon in their reading of the Chinese language. Frequency statistics are extremely useful not only for linguists conducting corpus-based research, but also for teachers in preparing textbooks and other pedagogical materials, and in designing tests that reflect vocabulary and linguistic patterns that the students regularly encounter.

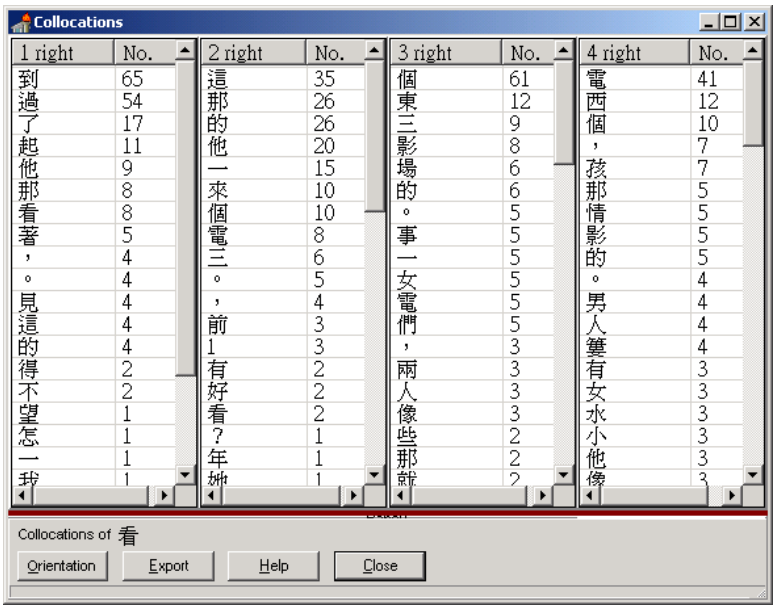
Figure 10. A full concordance of the Pear Stories (character-spaced texts).



The word highlighted in the Headword window in Figure 10 is 看 *kàn* 'to look,' with 209 hits. These 209 tokens are, in turn, displayed in the Context window to the right, with sorting there by the string that immediately follows the keyword. Scrolling down the Context window would enable the user to see what Chinese characters come after the keyword and how frequently those characters occur in that slot in the proximity of, or adjacent to, the keyword.

With the character-spaced e-texts for the Pear Stories, a study of collocations is also possible—a second major advantage of spaced e-texts over non-spaced e-texts. The keyword 看 in Figure 10, for example, has right collocates that are shown in Figure 11. While both left and right collocates are generated by the concordancer, Figure 11 focuses on the verb's right collocates, which were automatically sorted by frequency by the concordancing program. Of particular importance in this case are the right collocates that immediately follow the keyword. As shown there, 看 often co-occurs with 到 *dào* (and much less so with 見 *jiàn*) to form resultative compounds in this corpus. The verb also occurs frequently with the aspect markers 過 *guo*, 了 *le*, and 著 *zhe*. There are also several instances of reduplication of the verb.

Figure 11. Right collocates of 看 in the Pear Stories (character-spaced texts).



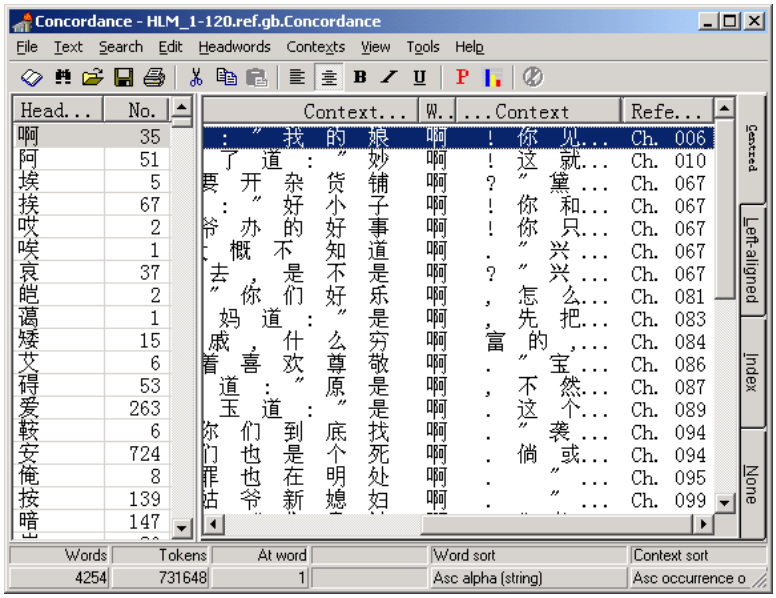
Another set of examples is given in Figures 12 through 15 using a literary source. Figure 12 displays a full concordance of a GB-encoded e-text of the novel, *Hongloumeng* 红楼梦 (Dream of the Red Chamber). The concordance of the 120 chapters was generated in about five minutes.[7] The highlighted keyword

[7] The e-text was from *Wenlin* 2.5, where the source of the original e-text was noted as follows: 'This version of 红楼梦 Hóng Lóu Mèng (Dream of the Red Chamber) was downloaded in October, 1998 from the website of 李晓渝 Lǐ Xiǎoyú, also known as Grand Master of the Great Empire of China <www.ifcss.org/xiaoyu-collection>.' My thanks to Thomas Chan for preparing the character-spaced e-text of *Hongloumeng*, which he initially used for a homework assignment for my Winter Quarter 2001 graduate Chinese linguistics seminar, 'Databases and Corpora for Chinese Linguistic Research.' To the character-spaced e-text, I have added the coding for chapters, for input into the *Concordance* program.

The full concordance took about five minutes. (See footnote 5 for factors that can affect the time that it takes to create a concordance.) For an interesting comparison, see the 'readme' file that accompanied the DOS-based CKWIC concordancer that David Steelman developed in November 1993, a program that makes a KWIC concordance for one character at a time from Chinese e-texts. He wrote in his 'readme' file zipped with the program, 'I am not a programmer so the code for this routine will undoubted [sic] be quite laughable to experienced programmers. Since nothing seems to be available at present I submit this in the hopes that someone will do a better job of it. There are some rather sophisticated programs available like TACT, Micro OCP, and the Longman Mini Concordance for handling one byte code. It would be nice to have something like that for Chinese. Being in need of a Chi-

is the particle, 啊 *ā* (also in Tones 2, 3, 4, and 0 'neutral tone'). The Context window is unsorted; hence, the default ordering is that of occurrence in the source e-text, containing all 120 chapters of the novel. Also displayed is the reference column showing the source chapter for a particular context line. (Coding of chapters in the source e-text was done before making the concordance.)

Figure 12. Full concordance of *Hongloumeng*: 'Alphabetic' sorting (by string in ascending order), with first keyword 啊 highlighted.



Interestingly, as shown in the figure, of the 35 tokens of the particle in the e-text, 7 (or 20%) of them occurred in the first 80 chapters; and of those 7 tokens, 5 occur in Chapter 67 alone. The remaining 28 (80%) are found in the last 40 chapters of the novel. Whether the skewed distribution of 啊 is but one of the various clues that the novel is the work of two different authors is a topic that can be ex-

---

nese concordancer, I fuddled through and made this. It's slow. It's primitive. But until something more sophisticated is made available, it does crank out the necessary concordance. For a full concordance of all characters, I added a character generator which made a concordance of each of the characters in the character set and sorted using BIGSORT (available from Simtel 20 as I recall) and appended to a file. It took something like five days to build a complete concordance for the Hong Lou Meng.' As one can see from the quote, we have come a long way since November 1993 when it took five days to generate a full concordance, compared to five minutes to accomplish the same task in the current Windows program. Some years from now, it will only take five seconds.

plored further, together with a study of other words in the novel. Concordancers can take much of the tedium out of manual searching for similar purposes, such as in Yu's 1996 study of interrogatives in the novel, 儒林外史 *Rulin Waishi* (The Scholars), where Yu finds differences between the first thirty-two chapters and the remaining chapters with respect to distribution and use of interrogatives in the novel that suggest the possibility of dual, or multiple, authorship.

The keywords in the Headword window in Figure 12 are sorted 'alphabetically' (by string, in ascending order) based on the encoding system, in this case GB-encoding. As a result, the sorting observes the ordering of Chinese characters based on Pinyin romanization, including ordering based on the four tones in Mandarin Chinese. The Pinyin ordering holds for close to four thousand frequently-occurring Chinese characters, with the remaining characters whose Pinyin romanization begins with 'z,' as can be seen in Figure 13, with highlighting placed on the six-token character, 乜 *miē*, as in 乜斜 *miēxie* 'look askance, squint' (also pronounced Niè as a surname).[8]

A quick concordance of the e-text for studying only certain words can also be conducted, as shown in Figure 14, comparing the coverbs, 把 *bǎ* and 将 *jiāng*, with default sorting in the Context window. (The quick concordance took just 19 seconds.) The two keywords have roughly the same frequency of occurrence in the e-text: 1142 tokens of 把 and 1160 tokens of 将.

However, the frequency data does not tell the same story concerning the distribution pattern of 把 versus that of 将. They are quite dissimilar, as can be seen in the comparison of the right collocates of these two keywords in Figure 15.

[8] Observe that this Chinese character, although rarely used in standard Chinese, is among the most commonly-used characters in the inventory of vernacular Cantonese characters. As T. Chan (2001: 62) observes, it is a phonetic loan in Cantonese: '*mat*^{1} (what), also pronounced *me*^{1} as a contraction, is actually a phonetic loan of *me*^{2} 乜 (to squint) based on the latter pronunciation, which differs in the tone, *yinshang* 陰上 (tone #2) rather than *yinping* 陰平 (tone #1).'

Figure 13. ‘Alphabetic’ sorting of the *Hongloumeng*,with 乜 highlighted.

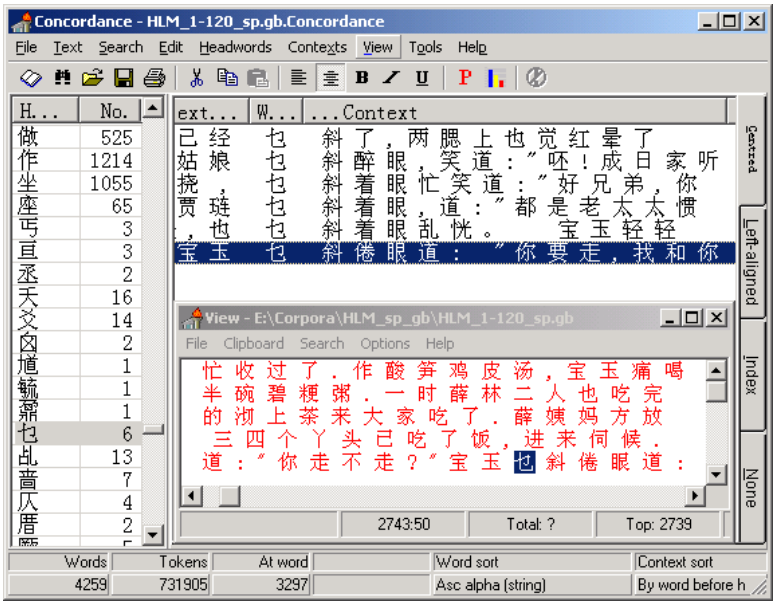


Figure 14. Concordance of 把 and 将 in the *Hongloumeng*.

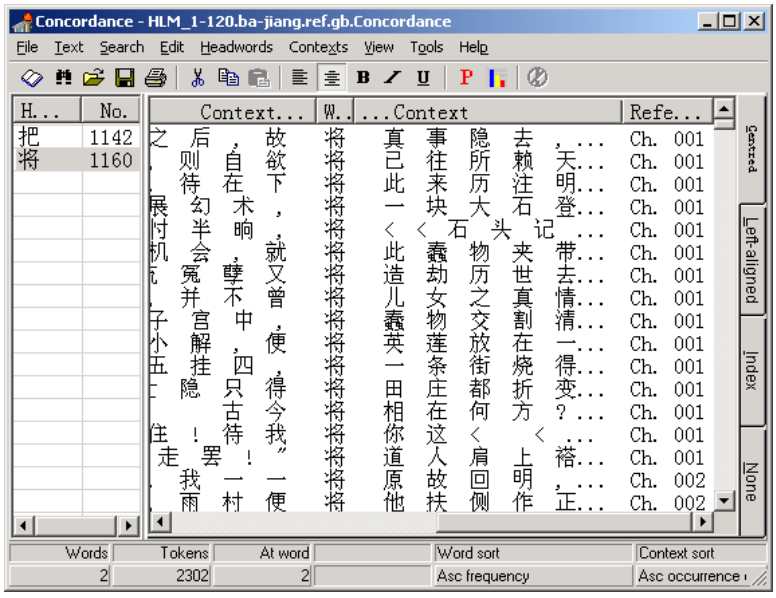
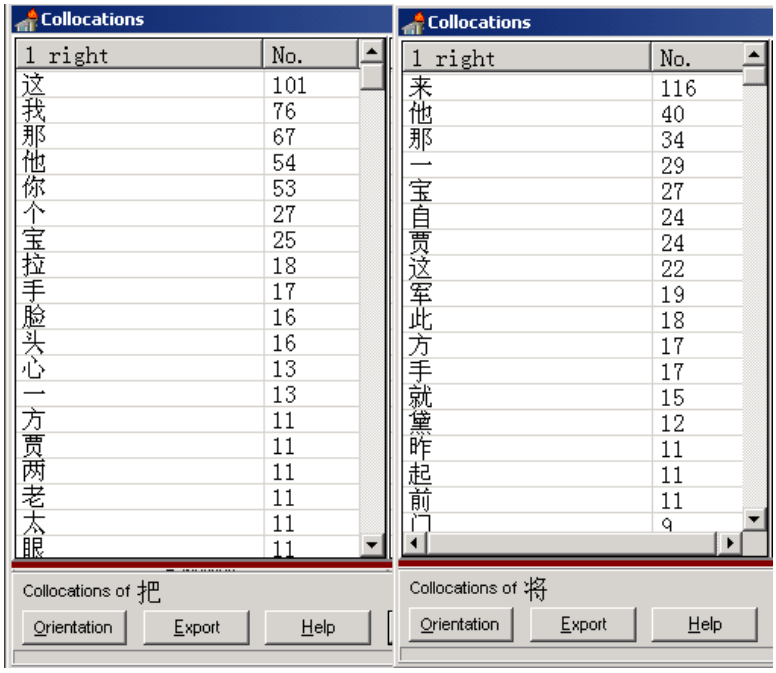


Figure 15. Right collocates of 把 and 将 in the *Hongloumeng*.



A closer scrutiny of the results is shown in Table 1, with the seven most frequently-occurring collocates of one or the other keyword functioning as a coverb, and listed in order of frequency (descending order). The number of tokens is also given. Omitted in (1) is 宝 *bǎo* 'precious,' occurring 27 times with 将 (and 25 times with 把), as it is part of a given name, combining with 玉 *yù* 'jade,' in the name of the protagonist, 宝玉 Baoyu. Also omitted is the most frequently-occurring collocate of 将, namely, 来 *lái* 'come' (116 tokens), since 将 is not a coverb in that context; instead, 将 combines with 来 to form the compound, 将来 *jiānglái* 'future, in the future.' (Other compounds containing 将, such as 将近 *jiāngjìn* 'close to, nearby; almost,' are far less frequent in the corpus.) With respect to 把 and 将 used as coverbs, they frequently co-occur with demonstratives and pronouns. However, overall, the coverb, 把, occurs more frequently with demonstratives, 这 *zhè* 'this' and 那 *nà* 'that,' than does 将. At the same time, coverb 把 occurs much more frequently with the first- and second-person

pronouns, 我 *wǒ* 'I' and 你 *nǐ* 'you,' suggesting that 把 might occur more frequently in the dialogues in the novel. While more detailed study will still be needed, one can, nonetheless, infer from the distribution patterns shown here that in this novel, coverb 把 is used in a larger number of colloquial, spoken contexts than is the corresponding coverb, 将. A similar observation might also be made of the general classifier, 个 *ge*, when it occurs directly after the coverb 把.

1. Frequency statistics on the right collocates of the coverbs, 把 and 将.

	Right collocates of 把		Right collocates of 将	
(a)	这	101	这	22
(b)	我	76	我	5
(c)	那	67	那	34
(d)	他	54	他	40
(e)	你	53	你	7
(f)	一	13	一	29
(g)	个	27	个	1
	Total	391		138

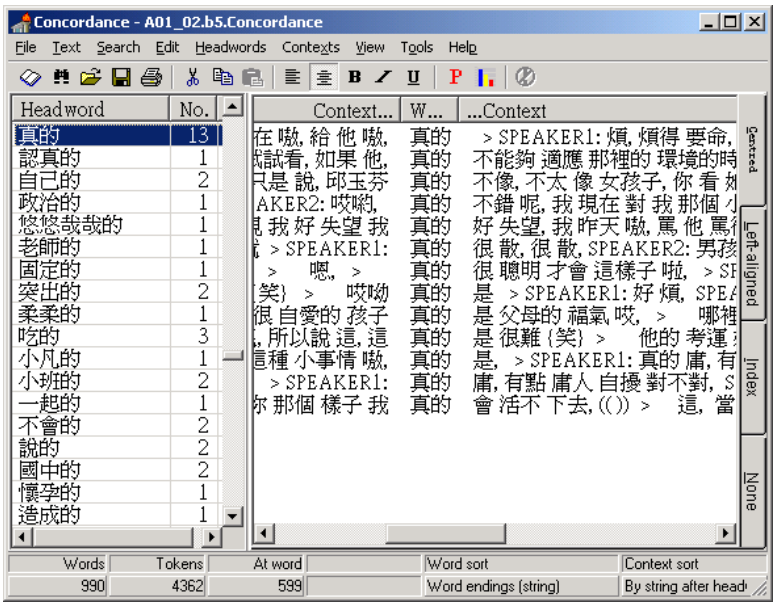
The preceding set of examples from a literary work as an e-corpus illustrates some directions that one could explore. Knowledge of collocations and their frequencies (in general, or for some specific corpus) would aid teachers in teaching vocabulary and the structure of the language, be it modern or some earlier stages of the language. At the same time, teachers who are familiar with concordancers and concordancing can guide their students to investigate the structure of the language and how it is used in different contexts, different registers, different genres, and different historical periods. Depending on the students' level of L2 acquisition, this can be accomplished through tasks that use either authentic texts serving as e-corpora, or pedagogically-prepared e-texts that contain vocabulary and sentence patterns that are familiar to the learners.

3.4. Using Word-Spaced Chinese E-Texts

Segmenting e-texts with spacing between Chinese characters is a very mechanical process, but the same is not true for segmentation of a linear sequence of Chinese characters into polysyllabic words separated by spaces. There is no

uniformly agreed-upon scheme for segmenting romanized Chinese text, for example, and there would be none for word-segmented e-texts. And in making a concordance, not all e-texts are segmented identically; it depends on what one is interested in searching for. In Duanmu et al.'s (1998) *Taiwanese Putonghua Speech and Transcripts,* for instance, the subordinative particle, 的 *de*, is treated as a suffix that is part of the word preceding it. In Figure 16, for example, of the 990 unique words in the A01-02 transcript of a dialogue, many are suffixed by 的. The thirteen-token keyword, 真的 *zhēnde* 'real, really, truly' is highlighted in the Headword window.

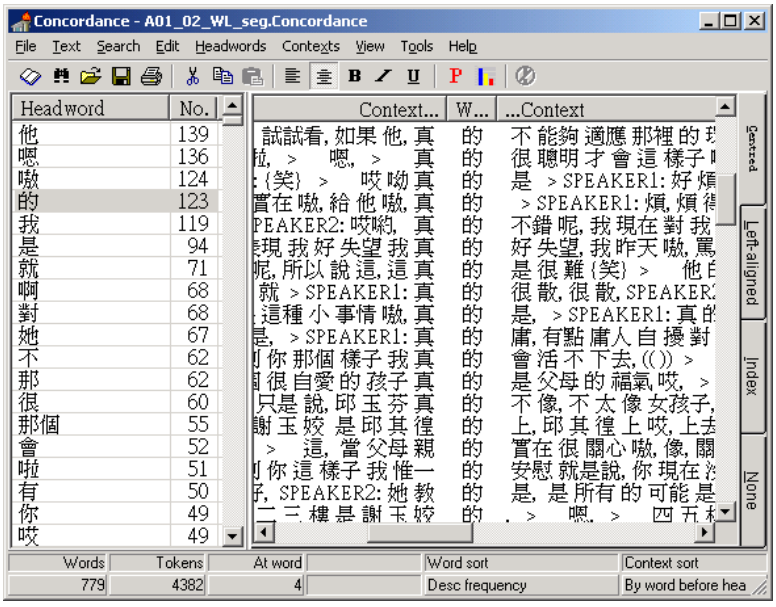
Figure 16. 的-suffixed words in a Taiwanese Putonghua transcript.



While it may be meaningful to treat 的 as part of the preceding word in romanized texts, treating each individual adjective, noun, and pronoun suffixed by 的 as a separate, unique word in a concordance may be less useful. It would be similar to treating every English word written with an apostrophe-s as a separate word from one lacking the apostrophe-s and entering both into the lexicon. The larger the e-corpus, the greater would be the proliferation of words containing 的 as separate entries. Even in this corpus alone, in addition to two occurrences of 自己的 *zìjǐde* 'one's own,' for instance, there are ten tokens of 自己 *zìjǐ* 'one-

self.' Multiply that for the other 的-suffixed words, and the distributional patterns that one observes from the results of such a full concordance would be very different from one in which commonly-occurring suffixes such as 的, and other grammatical markers, including utterance particles, are separated from the preceding word. (See also Wu (1998) on segmentation of Chinese.) Thus, we here further segmented that e-text using the segmenter bundled with *Wenlin*, and then made another full concordance. This yielded a total of 783 unique words, as displayed in Figure 17, with 123 occurrences of 的 in the corpus. The Context window is sorted by the string before the keyword, with the display of the 17 occurrences of 真 *zhēn* 'real' before 的.

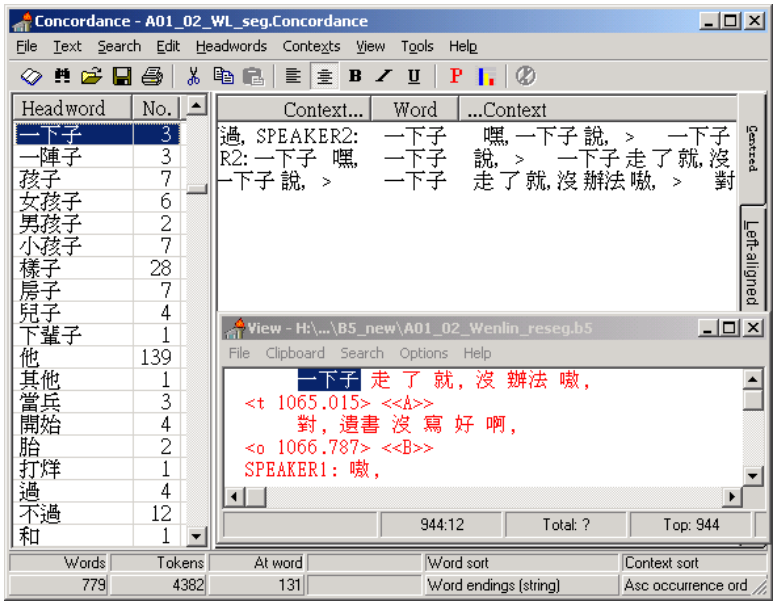
Figure 17. 的 as a keyword in a Taiwanese Putonghua transcript (word-spaced e-text further segmented using *Wenlin* 2.5).



A full concordance of word-spaced e-texts provides an efficient means to study vocabulary and morphological structure. A number of nouns contain the nominalizing suffix, 子 *zi*, for example, as can be seen in Figure 18, using the same Taiwanese Putonghua corpus. (Note also that the Taiwanese Putonghua corpus consists of transcripts of spoken corpora with speakers recorded in Taiwan; hence, as one might expect, there are no r-suffixed forms in the transcript.)

Without exception, all ten instances of 兒 *ér* 'child' in the corpus are either in the word for 'son' 兒子 *érzi*, or in the word for 'daughter,' 女兒 *nǚ'ér*.

Figure 18. Nouns containing 子 in a Taiwanese Putonghua transcript (word-spaced e-text).



Full concordancing of word-segmented e-texts also allows the user to browse through the keywords and spot near-synonyns and study how they differ in usage. For example, a student might explore how 向來 differs from 從來 in usage by studying the actual contexts in which these words occur, rather than be given a few isolated sentences as examples. Or s/he might explore how the adverb, 馬上 mǎshàng 'at once, right away,' differs from the near-synonymous adverb, 立刻 lìkè 'immediately, at once.'

As we see from the discussion in this subsection, word-segmented e-texts that closely resemble entries in a word dictionary (詞典) will likely be more useful for concordancing in an L2 environment than word-segmented e-texts that reflect word-spaced romanized texts. Besides being able to study near-synonyms and morphological structures in the lexicon in such concordances, word-frequency counts could also be conducted more easily. One can also keep in

mind that further patterns of language use can be obtained via regular expression searches to retrieve subsets of a string, such as determining the distribution pattern of the suffix 子 *zi* versus that of 兒 *er* in a corpus.

3.5. Using POS-Tagged Chinese E-Texts

Large e-corpora (one million words or more) that are tagged for parts of speech (i.e., POS-tagged) have long been available for English, with 1961 marking the beginning of the compilation of the first machine-readable corpus for linguistic research, the 'Brown Corpus' (i.e., Brown University Standard Corpus of Present-Day American English), completed in 1964 with over one million words (Kennedy 1998:23-24). General as well as specialized e-corpora exist for English, including those for the spoken register—with corpora from spontaneous as well as prepared speech—and the written register—consisting of written texts. Just as there is a lag in using concordancers for Chinese, POS-tagged e-corpora for Chinese are very scarce and typically are not readily available for the general user. Exceptional is the Sinica Corpus at Academia Sinica, Taiwan, available on the World Wide Web with an online search engine. The Sinica Corpus is open to the public for conducting searches with results in KWIC display format, as shown in Figure 19, where the search was for 條 *tiáo* as a classifier, thus excluding polysyllabic words containing 條, such as 條件 *tiáojiàn* 'condition, term,' and 條約 *tiáoyuē* 'treaty, pact.' The search yielded a total of 1,724 hits, and is exhaustive, since the maximum number of hits for any search is 2,000.

Display of the POS-tagging in the corpus was suppressed in Figure 19 as the default KWIC display format. Selecting viewing of the tags results in the display in Figure 20, which shows a corpus with word-segmentation, though without spacing between words per se. The POS-tagging is placed immediately after the word, such that the classifier 條 is tagged as follows: 條(Nf). Other classifiers are similarly separated out and tagged with 'Nf.' Observe also that the previously-mentioned subordinative suffix, 的, is also separately tagged as: 的(DE), with DE-tagging also used for the literary counterpart, 之 *zhī*, as can be seen in the web page of concordancing results, a small portion of which is shown in Figure 20. For corpus linguistic research as well as for concordancing in a language-learning environment, morphologically-based 'word' separation or POS-tagging would be most useful for multiple purposes. Alternatively, for the greatest flexibility, the same e-corpus can have different versions, depending on needs and purposes.

Figure 19. KWIC display of an online search for 條 in the Sinica Corpus.

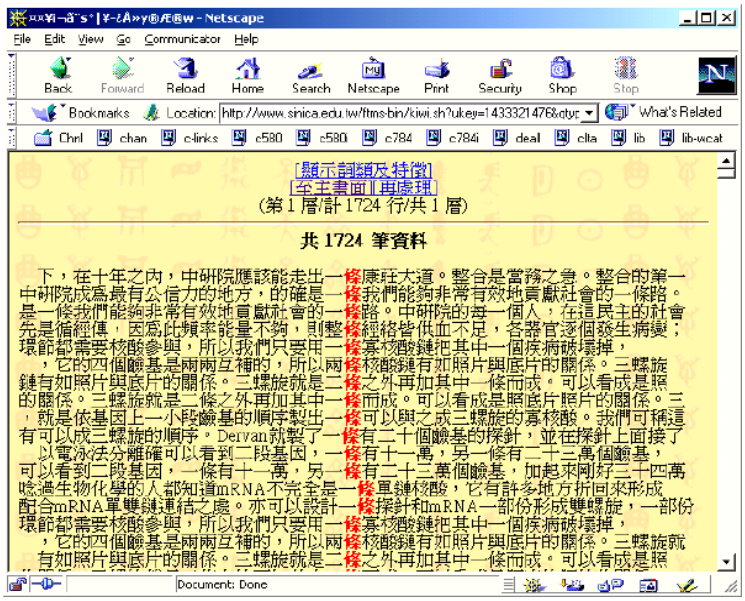
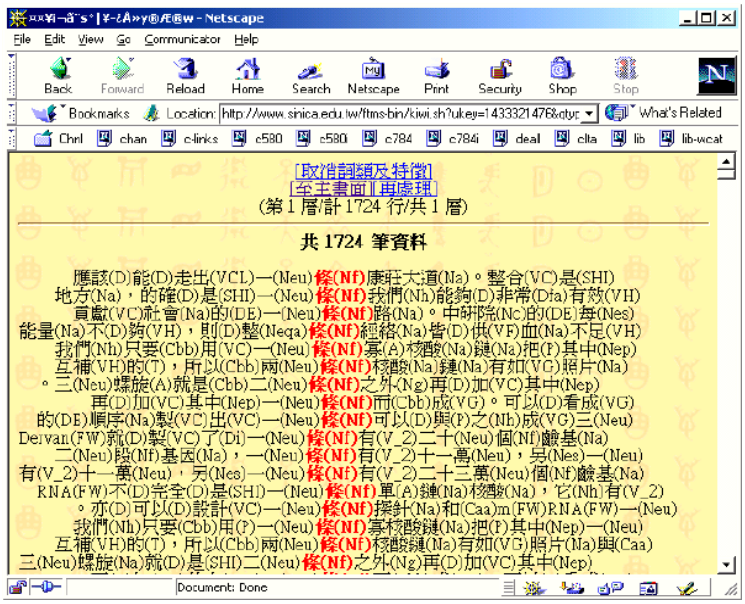
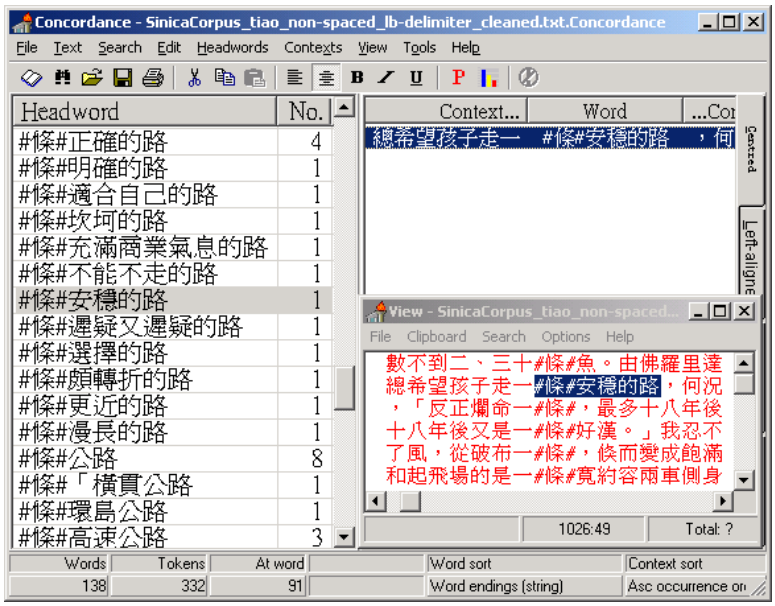


Figure 20. POS-tagged KWIC display of an online search for 條 in the Sinica Corpus.



In Figures 19 and 20, observe that the 1,724 tokens of the classifier, 條, are not sorted in any way. Hence, it would be practical to have the output of the on-line search serve as input for a concordancer in order to conduct a more detailed and systematic study of the 條 corpus. For example, the output in Figure 19 was prepared for concordancing by replacing the HTML-tagging around the keyword with the pound-sign (#) to identify those 1,724 keywords, along with performing a few other small, formatting tasks. With the new file as the source e-text, a regular expression search was conducted on '#條#', and the results were then sorted. The results show that a significant portion of the nouns associated with 條 in the corpus involve words that mean 'path,' 'road,' 'street,' 'lane,' and 'way,' and the most frequent among them is 路 *lù* 'road, path, way.'

Figure 21. Regular expression search for 條 and 路 in the Sinica Corpus.



Focussing on 路 and limiting the search to those phrases in the corpus that contain 條 and 路, a second search was conducted to retrieve all those lines with up to 10, and then up to 26, 'characters' between them.[9] This yielded some 300

[9] I first tried 10 characters, and then 20, and lastly 26 characters between 條 and 路 to make sure that the results obtained catch all relevant lines of the original 1,724 hits. That also meant that the results

or so hits, of which 100 or more are unique 'words,' as shown in Figure 21.[10] Studying this smaller concordance, one quickly notices that phrases with 路 in the corpus are being used both metaphorically and non-metaphorically to refer to 'road,' an observation that can also be made for English (and likely for many other languages). That is, from 路 to denote the physical entity, 'road,' on which one walks physically, via metaphorical extension, 路 is also used in more abstract senses, including instances where time is viewed in spatial terms, with a road extending in space becoming a projection in time. In the highlighted example, for instance, the speaker expresses his/her thoughts that those who are in the position of being family members inevitably hope that their children will travel a smooth and steady road (or journey in life through time).[11] Studying semantic extensions in word usage—including the investigation of literal and figurative uses of language, in the metaphorical extension from concrete to abstract—is yet another use of concordancers and concordances, is a means to enriching one's knowledge not only of the morphological and grammatical structures of the language, but also of that language's semantic properties.

It should be noted that concordancers are not the only tools that one might use with the 條 corpus. The pound-sign (#) flanking 條 can also serve as a delimiter separating each line of hits into three, ordered parts: context before the keyword, the keyword itself (條), and the context after the keyword. The text could be imported by a spreadsheet program (such as Microsoft's spreadsheet *Excel*) and manipulated from there, including adding columns for finer sorting, as shown in Figure 22.

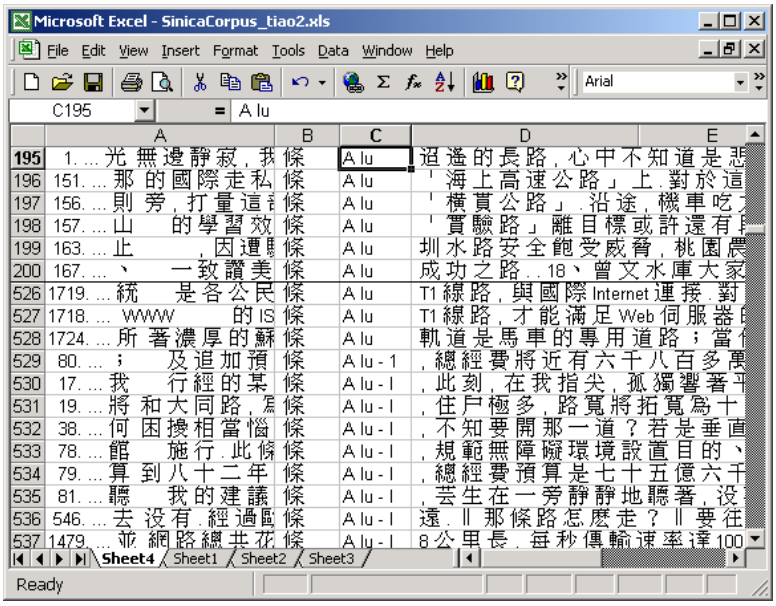
---

of this second search needed to be scanned manually to delete at least the obvious cases of irrelevant hits, such as hits where 條 and 路 occurred in separate phrases, and might even be separated by some punctuation mark. The syntax for the third and final regular expression search in the concordancer was the following, between the single quotation marks: '#條#.{0,26}路'.

[10] While the figure shows 138 unique 'words' with a total of 332 tokens, this can only be a rough figure, as the results were quickly prepared to use as an example. In the regular expression search conducted for this example, observe that the search results would automatically exclude those cases in the corpus where the noun precedes the classifier in the sentence. See, for example, the next discussion involving Figure 22.

[11] See Chan (1989), for example, for a brief discussion on the use of spatial terms and their metaphorical extensions to more abstract, temporal concepts of past and future.

Figure 22. The 條 corpus imported to a spreadsheet program.

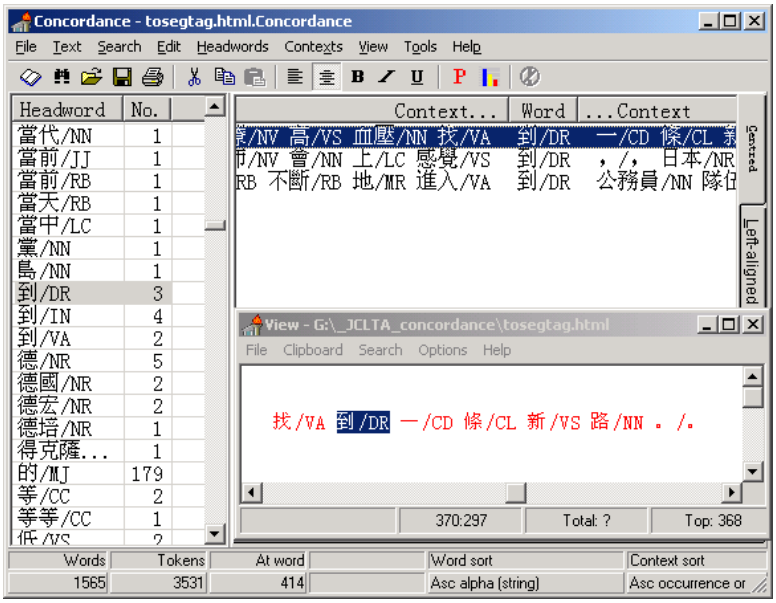


In the spreadsheet in Figure 22, a column ‘C’ was added after the column containing the keyword, 條; upper-case letters were used for major grouping of the different types of nouns that were associated with 條 in the corpus, and then further subgrouping was made within the major groups. After seeing that words with the meaning of ‘road,’ ‘street,’ ‘path,’ and ‘way’ were by far the most frequent collocates of 條, the letter ‘A’ was assigned to this group of nouns. For the context lines where the collocate is 路 as the head noun, these rows in the spreadsheet are further sub-classified with ‘lu’ added thus: ‘A lu.’ Typically, the collocates of 條 occur after the classifier, but there are occasions when the noun associated with the classifier precedes it. On those rarer occasions, a subclassification is made by separating these out and adding the subgrouping of ‘l’ for ‘left’; thus, these are subgrouped as ‘A lu – l’. (An alternative, of course, would be to treat these groupings and subgroupings as separate columns, C, D, and E, but the subgrouping given here suffices to illustrate.) In the corpus, there are over 300 instances where 路 serves as the noun associated with 條. This is reflected in the sorted spreadsheet, with sorting (in ascending order) first on Column C and then on Column D. Line 195 begins the set of tokens of 條 where 路 is the head

noun, and line 537 ends that set. The pane in the spreadsheet was ‘frozen’ to display the first six lines (i.e., lines 195-200, generated by the spreadsheet), in which Column C contains ‘A lu,’ and the last twelve lines (lines 526-537) of the spreadsheet, where Column C contains ‘A lu’ or ‘A lu – l.’

In this example, a spreadsheet program has been used. For a full-scope project for further investigation and querying, the file could also be exported to a database program with fields corresponding to the columns created in the spreadsheet. Observe that concordancers are, in many ways, simply another type of tool for querying a database, with the database here simply the source e-texts. Parts-of-speech tags in an e-text could, in a database program, be assigned to a field separate from the words themselves.

Figure 23. The 100 Sentence Corpus and 到 entries.



POS-tagged corpora are the most useful for corpus-linguistic research, so that time is not spent in further sorting of words that belong to more than one syntactic category, for example. This is illustrated in Figure 23, with a full concordance made of Okurowski and John Kovarik’s (1998) online POS-tagged corpus, *The 100 Sentence Corpus* (part of the Chinese Treebank Project). As one can see, there are three separate entries for 到 *dào* as a result of POS-tagging,

namely, as a 'directional complement' (DR), a 'preposition' (or 'coverb') (IN), and as a verb that is 'active' (versus 'stative') and hence an 'active verb' (VA). There are three occurrences of 到 as a 'directional complement.' In the Context window, the first one is highlighted. There, displayed in the View window at the bottom of the figure, is the keyword located in the source e-text, together with its left and right contexts. Only a small part of a fairly long sentence can be seen, namely, 找到一条新路 *zhǎodào yītiáo xīn lù* 'found a new road (route, method).'

Observe that in the corpus, the slash (/) is used to separate the word from the part-of-speech tag, and punctuation marks are likewise tagged. In the segment shown in the source e-text in Figure 23, which is part of Sentence 35 in the corpus, 找 *zhǎo* 'to seek' is tagged as an action verb (VA), and 到 *dào* as a directional complement (DR)—a category that actually includes resultative complements, such as 到. The resultative verb, 找到 *zhǎodào* 'seek-arrive – find' is then followed in sequence by: 一 *yī* 'one' a numeral (CD), 条 tiáo a classifier (CL), 新 *xīn* 'new' and lastly, 路 *lù* 'road' a common noun (NN), in contradistinction to a proper noun (NR).[12]

Until POS-tagged Chinese corpora are more readily available for public use, however, it is unlikely that many Chinese-language teachers and their students will be making use of such corpora in the very near future for concordancing and studying the structure of the language.[13] Note that one could, of course, manually POS-tag only those cases that one is investigating, such as tagging only those tokens in an e-text where 把 and 将 are coverbs, and retrieving those cases in a more refined concordance search (as opposed to the non-POS-tagged concordance, illustrated in Figure 14).

In this section of the paper, we have explored the different types of corpora for Chinese concordancing based on whether or not there are white spaces in the e-text, and whether or not the corpus has POS-tagging. Besides tagging of parts of speech, corpora could have annotations for prosodic structure, morphological structure, syntactic structure, pragmatic and discourse information, and so forth. The types of formatting and annotations will determine, in many ways, the kinds of research questions that can be addressed and how easily the information can

[12] See the list of syntactic tags presented under 'Segmentation Guidelines' at the Chinese Treebank Project Home Page, <umiacs.umd.edu/labs/CLIP/forest.html>.
[13] Academia Sinica, for example, makes available a segmentation and tagging program that runs under (Traditional Chinese) Windows. URL: <godel.iis.sinica.edu.tw/CKIP/>.

be retrieved using a concordancing program. In the next section, we will discuss the use of concordancing in language teaching and linguistic research, together with some slightly more elaborative examples.

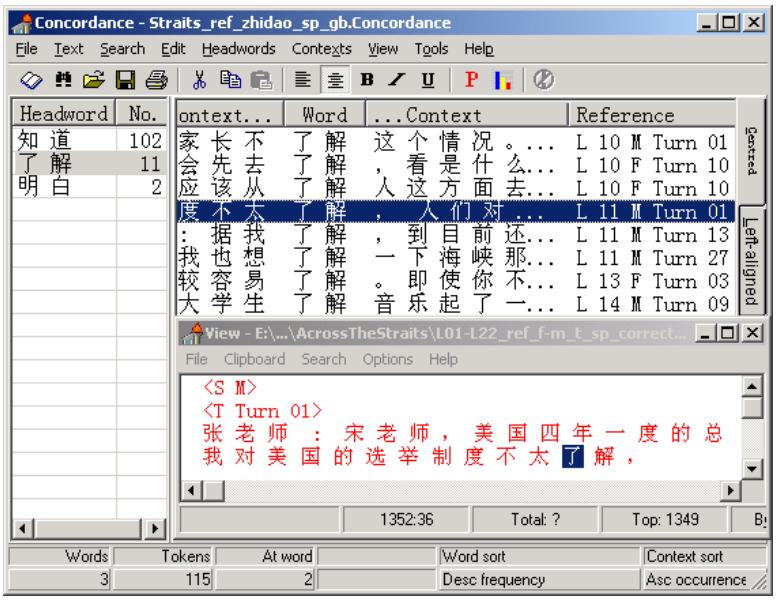
4. Concordancing and Chinese Language Teaching

The different types of e-texts and what kinds of searches one can do were discussed and illustrated in section 3. As shown in the examples, one can study the morphological structure of the lexicon, collocations, distributional patterns of near-synonyms, and so forth. For students learning to read and write Chinese, concordances, with their visual mode of presentation, serve as a valuable aid in providing plenty of examples of word usage in "real," non-grammarbook contexts. Of particular value then are concordances of e-texts of literary works, scholarly writings, newspaper articles, as well as writings in less formal registers, as a means to supplement grammatical explanations and examples accompanying vocabulary lists in textbooks, as well as reference sources such as Mickel (1999) and Teng (1996). In the illustrations in section 3, some of the sources for e-texts are transcripts of *spoken* corpora—both oral narratives and dialogues—that show how concordances for Chinese might range in register, from written to spoken, from more formal to less formal. Discriminating selection of the source e-texts, then, enables a student to compare, say, a formal written piece vis-a-vis a piece reflective of the spoken language, and thereby to more readily grasp distinctions between the spoken and written registers, between the formal and the less formal.

Pedagogically-prepared language-teaching materials with vocabulary and grammatical patterns introduced in each lesson are also prime candidates for concordancing and reviewing of vocabulary and grammatical patterns. For example, in the dialogues in the Bai et al. (1998a, b) miniscripts, *Across the Straits,* the adverb, 往往 *wǎngwǎng* 'often, frequently,' is introduced in Lesson 3, with three occurrences of it in the dialogue in that lesson; it recurs in Lesson 15 and again in Lesson 22, resulting in a total of five instances in the set of twenty-two lessons, with each lesson containing one dialogue between two teachers, Sung Laoshi (female teacher from Taiwan) and Zhang Laoshi (male teacher from mainland China). At the same time, a near-synonym, the adverb, 常常 *chángcháng* 'frequently, often, usually,' can also be found in those dialogues, with a total of 12 occurrences. The students are assumed to already know 常常, since the dialogues were designed for developing advanced listening skills in Chinese. With a concordance of just those two words and all instances in those twenty-two chapters,

students can study the occurrences of those two adverbs during the course, as one more class activity, or for review at the end.

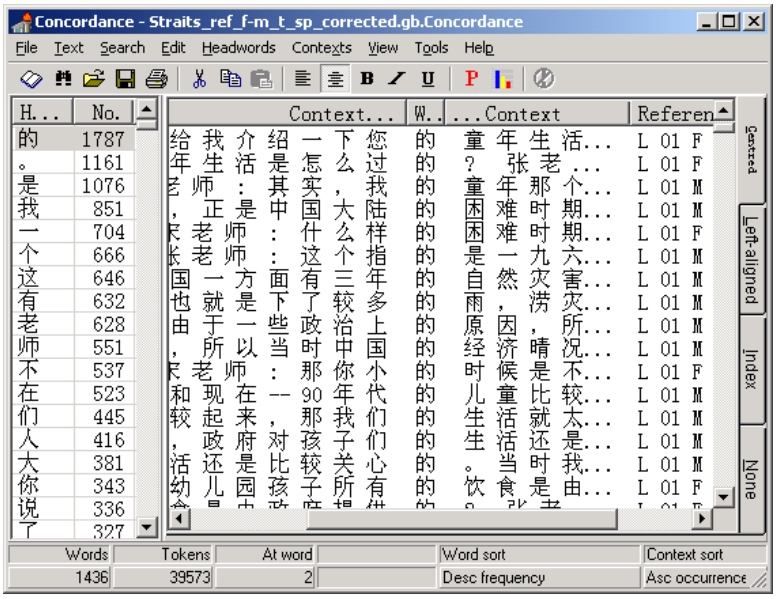
Figure 24. 明白, 知道, and 了解 in *Across the Straits*.



Similarly, one finds the near-synonymous verbs, 明白 *míngbai* 'understand, realize, know,' 知道 *zhīdao* 'know, realize,' and 了解 *liǎojiě* 'understand, comprehend.' As shown in Figure 24, of these three verbs, 知道 occurs most frequently (102 tokens) and 明白 the least (2 tokens, of which one is used adverbially) in the twenty-two dialogues. With the convenience of concordances, students can study the contexts in which these verbs are used in those dialogues. In Figure 25, 'References' are displayed to identify: (1) the lesson containing the token, (2) the speaker ('F' for female teacher, Sung Laoshi, and 'M' for male teacher, Zhang Laoshi), and (3) sequencing of speaker turns within a dialogue. That is, the source e-text was marked up to include information on lesson numbers, speakers, and turns, which the concordancer recognizes and processes. Indexing of lessons gives the teacher the flexibility to select words from specific lessons or from the entire book.[14]

[14] The transcripts were from an earlier, prepublication version of the dialogues that form the mini-scripts for *Across the Straits*. Thanks go to Bai Jianhua for providing me with a set of *Xin Tianma* files

Figure 25. Full concordance of *Across the Straits* (character-spaced e-text).



A full concordance of the lessons in a textbook would also be extremely useful, whether the e-text is word-spaced or character-spaced. Figure 25 illustrates a character-spaced, full concordance of the twenty-two dialogues in *Across the Straits*, with sorting of the keywords by frequency in descending order, and highlighting of the top keyword, the subordinative particle, 的 *de*. The concordance shown retains as 'keywords' clause- and sentence-final punctuation marks as well as two English words that are in the source e-text, along with some Arabic numerals, so that the total count of different (simplified) Chinese characters would be slightly fewer than the 1,436 keywords ('Words') displayed at the bottom left corner of the figure. Roughly speaking, a student who is familiar with every Chinese character in that set of 22 miniscripts would know over 1,400 Chinese characters. And the more frequent the character, the better s/he would

---

of the transcripts, which were converted to GB-encoding and combined into one file, with spacing and reference coding for lesson number, speaker, and turn number added to create the source e-text for this example. According to the authors (Bai et al. (1998a,b), *Across the Straits* may be used by itself or in conjunction with any of the following three textbooks published by Cheng & Tsui: *Taiwan Today, Beyond the Basics,* and *A New Text for a Modern China*, which was formerly *A Chinese Text for a Changing China.* (*Xin Tianma* is a DOS-based word-processing program from Asia Communications, and is still bundled with their current software to enable exporting of files created in their program to GB/Big5-encoded text files. (URL: <www.cjkware.com/>)

(or should) know that Chinese character. For teachers, the frequency data would be extremely useful for determining what to test and review.

While punctuation marks may be omitted in making a concordance, they are, nonetheless, useful to retain in a full concordance, as they can help students learn how to use Chinese punctuation marks. At the same time, they are also interesting for examining the linguistic structure of a corpus. In the *Straits* corpus, for example, there are 1,416 sentence-final punctuation marks: 1,161 periods for declarative sentences, 233 question marks for interrogative sentences, and 22 exclamation marks for exclamatory sentences. Since the e-text also includes gender of the speaker, a study of the corpus can include an exploration of potential gender differences. For instance, despite the female speaker taking more turns, a count of the number of sentences showed that 820 (58%) of the utterances were produced by the male speaker, contrasting with 596 (42%) produced by the female speaker. In fact, a rough 60-40 percent split is also found with respect to syllable count, namely, the male speaker produced 20,368 syllables (58%) and the female speaker produced 14,778 syllables (42%). The results indicate that, while the female speaker had more speaking turns, she talked less per turn overall than her male colleague. Interestingly, the proportion of roughly 60-40 percent with respect to male versus female sentence production is also found in Chan's (1996:20) study of a Cantonese corpus based on a set of transcripts from a television series. There, out of a total of 3,657 sentences, 2,217 (61%) were produced by the males in the episodes, and only 1,440 (39%) were produced by the females in the same set of episodes.

Encountering authentic language materials, students discover that men and women handle language differently—as the dialogues of the *Straits* corpus demonstrate. A teacher might conduct concordanced e-text searches, then discuss the results with students; or s/he might guide the students in conducting their own empirical studies of, say, distributional patterns, to be followed by student analysis and interpretation thereof. Here, two examples will be presented: gender-linked differences with respect to sentence types, and the use of sentence-final particles. Regarding different sentence types, they are not equally distributed by gender in the corpus. As shown in (2), the female speaker produced more interrogative sentences then the male speaker; in fact, almost two-thirds of the interrogative sentences in the corpus were produced by the female speaker. Only 10% of the male speaker's sentences are interrogatives, compared to 25% for the female speaker.

2. Distribution of sentence types in the *Straits* corpus.

	Sentence Types	Male	Female	Total
(a)	Declarative Sentences	727 (88%)	434 (73%)	1161 (82%)
(b)	Interrogative Sentences	85 (10%)	148 (25%)	233 (16%)
(c)	Exclamatory Sentences	10 (1%)	12 (2%)	22 (2%)
	Total	822 (100%)	594 (100%)	1416 (100%)

3. Sentence-final particles (SFP's) in the *Straits* corpus.

	SFP's	Male	Female	Total
(a)	啊	5 (4 decl., 1 excl.)	17 (8 decl., 7 int., 2 excl.)	22
(b)	吧	4 (4 excl.)	11 (5 decl., 5 excl., 1 int.)	15
(c)	了	49 (48 decl., 1 excl.)	25 (19 decl., 3 excl., 3 int.)	74
(d)	嘛	0	1 (1 decl.)	1
(e)	吗	5 (5 int.)	10 (10 int.)	15
(f)	呢	16 (16 int.)	24 (24 int.)	40
	Total	79	88	167

Other gender-related differences can also be studied, such as distribution of following sentence-final particles (SFP's) 啊 *a*, 吧 *ba*, 了 *le*, 嘛 *ma*, 吗 *ma*, and 呢 *ne*, as given in (3). Note that 嘛 occurs thrice in the corpus (counting one case transcribed as 吗 after the subject of a full sentence), and each time it was pro-

duced by the female speaker. In addition, there is one occurrence of the utterance particle, 啦 la, in clause-final position at the end of Lesson 5, produced by the female speaker: '五点半了，该吃饭啦，我们走吧!' (It's half-past five LE, time to eat LA. Let's go BA!) Each clause ends in a particle, whereas the male speaker's response contains no particles: '好, 那么下次再聊, 再见!' (Good, then, (let's) chat next time. Good-bye!) Both particles 嘛 and 啦 reflect gender-linked use that shows up in the corpus. Equally interesting is the observation in the corpus of the greater production of 吗-particle questions used by the female speaker. If this reflects a more general pattern of gender preference in the use of 吗 in interrogatives, it has not, to this writer's knowledge, been reported in the literature.

Observe, also, that although the two speakers produced roughly the same number of sentence-final particles (79 for the male and 88 for the female), these SFP's are not similarly distributed. For the male, well over three-fifths (56 out of 79) of his SFP's occur in declarative sentences, while just over one-fourth (21 out of 79) occur in interrogative sentences. In contrast, for the female, less than two-fifths (33 out of 88) of her SFP's occur in declarative sentences, while, at the same time, half (45 out of 88) of her SFP's occur in interrogative sentences. The particle 啊, which is used to soften the tone of an utterance, is produced more than three times as often by the female speaker, including in questions seeking confirmation. The male speaker uses only 呢 and 吗 in interrogatives, with much stronger preference for 呢. A follow-up study might explore the use of the A-not-A interrogative pattern, including A-not-A tag questions.

Recall also that the male speaker produced more sentences. Hence, even though the actual total number of SFP's produced differs little between the two speakers, the same difference shows up with respect to the proportion of sentences containing SFP's. The female speaker not only produces more actual SFP's, she also produces a higher proportion of SFP's in her speech than her male colleague: her 88 SFP's occur in approximately 15 percent of the sentences she produces, whereas the male speaker's 79 SFP's occur in only about 10 percent of the sentences he produces. The stereotypical impression that females use more sentence-final particles holds in this corpus, containing dialogues between two colleagues in the teaching profession. One additional variable that intersects with gender difference is a slight age difference between the two speakers, enough for there to be some social hierarchy at work, with the male speaker slightly older and hence slightly more senior. To what extent the differences in

their speech, especially the amount of talking, are due to factors involving social hierarchy, and to what extent to gender differences, need to be teased apart. In any event, the differences are there. At the same time, the stereotypical expectation that women be more polite may also be reflected in this set of dialogues, as only the female speaker uses the polite second-person pronoun, 您 nín.

Other, non-gender-linked, linguistic differences can also be observed. For instance, since the male speaker is from the mainland and grew up in Beijing (according to Lesson 13 on linguistic differences across the straits) and the female speaker is from Taiwan, there are some dialectal differences in their choice of vocabulary, as noted in Lesson 13. Other differences include the choice of classifiers: the speaker from Beijing uses 辆 liàng for 汽车 qìchē 'automobile' (and 摩托车 mótuōchē 'motorcycle, motorbike'), for example, whereas the speaker from Taiwan uses the classifier, 部 bù.

Furthermore, although one would not expect the female speaker to use r-suffixation, she did in fact do so on seven occasions in the corpus, while the male speaker used r-suffixation on twenty. In addition, given the male speaker's Beijing background, it is not surprising that he made a distinction between the inclusive first-person plural, 咱们 zánmen, and the exclusive first-person plural, 我们 wǒmen. All ten occurrences of 咱们 in the corpus were produced by the speaker from Beijing; the speaker from Taiwan used 我们 wǒmen consistently for both inclusive and exclusive first-person plural.

The examples presented here suggest areas where a corpus-based study of the language-learning materials can be conducted based on concordancing results. Teachers can use concordancers for course preparation prior to class, and can present examples derived from the concordances. They can also explore along with their students by conducting concordancing demonstrations in class, using the various course materials. And, lastly, there can be student-initiated, teacher-directed projects in using concordancers to study collocations, distributional patterns, and much more. And though concordancers are simply tools that teachers can bring to the language-teaching and language-learning environment, they do make possible the exploration of language and its texts in ways that would prove altogether too tedious and time-consuming via manual searching.

In the next section, we will briefly discuss sources of Chinese e-texts, including scanning as needed.

5. Sources and Preparation of Chinese E-Texts

Many Chinese e-texts, plain or containing HTML-tagging, can be freely downloaded from the Web. These include e-news items, e-magazine and e-journal articles, essays, short stories, novels, and any other Big5- or GB-encoded material that can be viewed on the Web. There are also bilingual Chinese-English e-texts of classical works online.[15] And a growing number of web pages are also available online for language learners, designed for specific levels, as illustrated in the following three websites:

4. Websites for e-texts for language learners

(a) Website: Reading: Progressive & General Readings
Sponsor: Chinese Language Program, U. of Southern California
Levels: Beginning to advanced levels
URL: www.usc.edu/dept/ealc/chinese/newweb/reading_page.htm

(b) Website: The Chinese Reading World
Sponsor: Chinese Reading Program, University of Virginia
Levels: Beginning and intermediate levels
URL: faculty.virginia.edu/cll/chinese_reading/

(c) Website: Archive of Chinese Teaching Materials
Sponsor: Chinese Language Program, Harvard University
Levels: Advanced level
URL: www.fas.harvard.edu/~clp/China/teach1.htm

In addition to the preceding, anything that an individual has prepared in digital format (particularly if it can be converted to plain Big5- or GB-encoded text files) might also serve as an electronic corpus for (Chinese) concordancing. For teachers, this includes course materials that might have been prepared using word-processing software. Not to be overlooked are electronic, searchable dictionaries, such as John DeFrancis' *ABC Dictionary* that is included in *Wenlin* (version 2.0 and higher), where search results can serve as source e-texts for concordancing software, as in one's analysis of morphological structure, of the multiple senses of words, and more.

[15] See, for example, my ChinaLinks web pages (URL: <deall.ohio-state.edu/chan.9/c-links.htm>) with links to numerous resources for digitized texts, as well as some links in my online Winter Quarter 2001 course page for Chinese corpus linguistics (URL: <deall.ohio-state.edu/chan.9/c889.htm>.)

Some e-corpora are also available commercially, such as those at the University of Pennsylvania's Linguistic Data Consortium (URL: <morph.ldc.upenn.edu>), including the Taiwanese Putonghua corpus, the Chinese TreeBank, and other Chinese spoken and written corpora.

For materials that are available in hardcopy format only, one solution would be to scan the texts using a scanner in combination with Chinese OCR software, and then proof-reading and correcting any errors introduced in the scanning process. Much of the archived e-texts online were scanned in this way, since retyping manually would be an even more time-consuming and labor-intensive job.

A few remarks are also needed concerning Unicode-encoded e-texts, such as UTF8-encoded web pages. With the exception of *Wenlin*, which is software for learning Chinese, the dedicated concordancing programs that this writer has tested cannot handle UTF8-encoded e-texts. Such e-texts would need to be converted to GB or Big5 encoding. Among the conversion software available is *NJStar Communicator'*s Universal Code Converter. Alternatively, one can save a file to Big5- or GB-encoding, in such Unicode software as *Wenlin, MS Word 97* (or higher), or NJStar's *Chinese Word Processor for Windows*. For Unicode capability, we shall need to wait for major updates to current concordancers.[16]

We conclude with a few remarks concerning segmentation. For character-space segmentation, one utility that has already been mentioned in section 3.3 is the Universal Code Converter. Unicode-compliant word-processing programs such as Microsoft's *MS Word 97* (or higher) can add manual line breaks to texts, as well as add spaces to non-spaced double-byte East Asian texts via its search-and-replace function. For word-spaced e-texts, in addition to the *Wenlin* software that is able to segment Chinese non-spaced text into (polysyllabic) words, Erik Petersen's online *Chinese Annotation Tool* and Chin-chuan Cheng's personal software, *CCLang*, which he developed for teaching Chinese dialectology and corpus linguistics, also has a utility for segmenting non-spaced Chinese text into (polysyllabic) words.

[16] Rob Watt (personal communication regarding in *Concordance* program) in the near future plans to explore some stopgap methods of handling UTF8-encoded e-texts. To fully be able to handle Unicode, the program would require a complete re-write, a task that will need more time, along with the right programming tools that support Unicode. In the meantime, Mike Scott (personal communication) indicated that his upcoming *WordSmith Tools 4.0* will be able to handle Unicode, and perhaps will be able to handle UTF8-encoded Chinese e-texts. How well *any* concordancer will be able to handle UTF8-encoded Chinese e-texts can only be determined after the program is available for testing.

6. Concordancing Programs for Chinese E-Texts

Several concordancing programs have been discussed in this paper, each of which was tested in English Windows 98 and Windows 2000. There have been earlier DOS-based programs, such as David Steelman's *CKWIC* program, developed in 1993 (see footnote 7) and Harry Yu's *Search* program for Big5-encoded e-texts, developed in the mid-to-late 1990's for (Traditional) Chinese Windows. The former is freeware, and the latter privately-owned and not available to the public.[17] Among the programs developed for Chinese is Tom Bishop's *Wenlin* program for learners of Chinese, which runs under DOS, Windows, and Mac operating systems. As stated in *Wenlin 2.0 User's Guide*, one can 'automatically search through any collection of Chinese documents to find examples of words and characters in context' (Tennenbaum and Bishop 1998:5). *CONCORD*, a concordance creation tool, is another program that works under DOS and Mac. Developed by Christian Wittern as part of the Zen KnowledgeBase project, the program automatically generates a concordance from a Chinese text file encoded in Big5 or JIS. The Summer Institute of Linguistics' *Conc* 1.80 program for the Mac can handle concordancing of Chinese. (*Conc* version 1.80 beta test version is freeware.)[18] Another Mac concordancing program, developed and privately-owned, is Patrick Moran's *GrabRec*, a line-oriented program that can do advanced searches. (He has other programs as well, such as *ChiFreq*, which counts the number of characters in a text.) Another setup that handles Chinese e-texts, along with English, Japanese, and French e-texts, is Chris Greaves' *ConcApp* concordancing programs, viz., his ConcApp Concordance Browser and Editor, a freeware program for Windows 95/98/NT, and his earlier *ConcApp Concordance Browser for Windows 3.1/95*.[19]

---

[17] Thanks go to Shou-hsin Teng for bringing the *Search* concordancing program to my attention in February, 1998.

[18] Thanks go to Olli Salmi for the information on SIL's *Conc* program for the Mac.

[19] Further online information on the concordancing programs cited here can be obtained at the following websites:

Besides the programs just cited,[20] there are also commercial concordancing programs intended for single-byte, Latin scripts that can also generate concordances of double-byte Chinese e-texts. These include Michael Scott's *Concord*, a 16-bit program packaged with his *WordSmith Tools 3.0,* which was developed for Windows 95 (and later, although for Chinese e-texts, the program runs best under Windows 95/98). *WordSmith Tools 4.0,* for Windows 98/2000/Me, which can handle Unicode, is not yet available (as of January 2002). Other programs include Michael Barlow's *MonoConc Pro 2.0* for Windows 3.1 and 95/98 and Rob Watt's *Concordance 3.0* for Windows 9x/NT/2000/Me. For Chinese concordancing in English Windows, these latter two programs, *MonoConc Pro 2.0* (which works best in Windows 9x) and *Concordance 3.0* (which works best in Windows 2000), have the most capabilities for concordancing tasks, with *Concordance 3.0* also capable of generating full concordances, as illustrated in the various figures shown here. However, as noted in section 3.3, currently, full concordances can only be made using spaced e-texts. Perhaps a Unicode-based, later generation of the program will be able to make full concordances of non-spaced e-texts.

---

i.

Software	**Website**
CKWIC	www.ifcss.org/links/software/cnapps/doscutil.html
ConcApp Concordancing Programs	vlc.polyu.edu.hk/pub/concapp/concapp.htm
Conc	http://www.sil.org/computing/conc/ (http://helios.unive.it/~pregadio/computing/conc/conc.html)
CONCORD	www.iijnet.or.jp/iriz/irizhtml/tools/concord.htm
Concordance 3.0	www.rjcw.freeserve.co.uk
GrabRec	---
MonoConc Pro 2.0	www.athel.com
Search	---
Wenlin 2.x/3.0	www.wenlin.com

[20] Other software programs and utilities cited in this paper include the following that can segment non-spaced Chinese e-texts: Universal Code Converter built into *NJStar Communicator* 2.x, Erik Petersen's *Chinese Annotation Tool*, Unicode-based *MS Word* (which can also perform double-byte character segmentation via its search-and-replace function), and Chin-chuan Cheng's *CCLang* program.

At this time, to the best of this writer's knowledge, there is no public-domain or commercial program that can make full concordances using non-spaced Chinese e-texts. To be able to generate such concordances, the concordancer needs to be able to treat Chinese characters as single, non-decomposable, multi-byte units, regardless of whether or not there is spacing between characters. Unicode-based word-processing programs, such as *MS Word 97* and higher, for example, are able to handle Chinese characters in this way.

For full concordances of non-spaced e-texts in which keywords are word-based and not simply character-based, an even more sophisticated program would be needed, namely, one that recognizes word-segmentation despite non-spacing of the text. Such concordancing programs for individual end-users are still somewhere beyond the horizon. In the meantime, in the Appendix, a comparison is given of some of the main features of four of the concordancing programs with respect to their functions and their ability to make concordances with Chinese e-texts in English Windows 2000. The four programs are: *Wenlin 2.5* (and 3.0 field test version)*, ConcApp 2.0, MonoConc Pro 2.0,* and *Concordance 3.0*.

7. Concluding Remarks

In this paper, an introduction to concordancing of Chinese is presented, with some simple illustrations on how concordancers can be used for making quantitative, empirical studies combined with analytical interpretations of the results. Concordancing is yet one further extension in the use of computer technology for both students and for teachers, and is another means to explore and learn the language. Teachers can use concordancing software for course-material preparations as well as for in-class concordancing tasks and teacher-guided, student-initiated concordancing projects.

---

ii.

Software	Website
NJStar Communicator 2.x	njstar.com
NJStar Chinese Word Processor for Windows	njstar.com
Chinese Annotation Tool	www-rohan.sdsu.edu/~chinese/annotate.html
MS Word 97 (or later)	microsoft.com
CCLang	---

Frequency data provide invaluable information for teachers on frequency of characters/words in a corpus, and allow for the exploration of which are the most frequently-occurring words in certain kinds of corpora versus some other kinds of corpora, such as written versus spoken, formal versus informal, standard versus regional varieties, gender differences in language use, modern versus other historical periods to study language change, figurative versus non-figurative use of language, typological differences across languages, and so forth. The study of collocations can extend into the study of semantic shift of the word under study. Furthermore, corpus-based studies provide frequency statistics on associations, such as certain nouns with certain classifiers, that can contribute to the research on determining central, or prototypical, members of a category.

With so much that is now digitized, depending on the student's language level, teachers can use either authentic materials or pedagogically-prepared e-texts as corpora. Searches can be made of specific words/phrases, or full concordances can be prepared for studying an entire corpus. Concordances offer a corpus-based, empirical approach to exploring lexical, morphological, syntactic, semantic, and discourse-level phenomena in the language.

References

All URL's to web pages in the World Wide Web (in both the references and footnotes) were last re-accessed on 19 January 2002.)

Bai, Jianhua, Juyu Sung, and Hesheng Zhang. 1998a. *Across the Straits: 22 Miniscripts for Developing Advanced Listening Skills in Chinese. Student Book* (Traditional Character Edition and/or Simplified Character Edition and audiocassette tapes.) Boston: Cheng & Tsui Company.

Bai, Jianhua, Juyu Sung, and Hesheng Zhang. 1998b. *Across the Straits: 22 Miniscripts for Developing Advanced Listening Skills in Chinese. Traditional & Simplified Characters Transcript.* Boston: Cheng & Tsui Company.

Biber, Douglas, Susan Conrad and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge, UK and New York: Cambridge University Press.

Chafe, Wallace L. ed. (1980) *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production.* Norwood, New Jersey: Ablex Publishing Corp.

Chan, Marjorie K.M. 1989. [Review] Shou-he Tian. 1989. A Guide to Proper Usage of Spoken Chinese. (Hong Kong: Chinese University Press.) *Journal of the Chinese Language Teachers Association* 24.3: 117-126.

Chan, Marjorie K.M. 1996. Gender-marked speech in Cantonese: the case of sentence-final particles *je* and *jek*. *Studies in the Linguistic Sciences* 26.1/2: 1-38.

Chan, Thomas. 2001. *Orthographic Change: Yue (Cantonese) Chinese Dialect Characters in the Nineteenth And Twentieth Centuries.* M.A. thesis, Ohio State University.

Duanmu, San, Gregory H. Wakefield, Yi-ping Hsu, Shan-ping Qui, and Guevara Rowena Cristina. 1998. *Taiwanese Putonghua Speech and Transcripts.* Produced by the Linguistic Data Consortium, University of Pennsylvania. (URL: <www.ldc.upenn.edu>.)

Erbaugh, Mary S. 1990. Mandarin oral narratives compared with English: The pear/guava stories. *Journal of the Chinese Language Teachers Association* 25.2:21-42.

Huang, Chu-Ren and Kathleen Ahrens. 1999. The function and category of *gei* in Mandarin ditransitive constructions. *Journal of Chinese Linguistics* 27.2: 1-26.

Hunston, Susan. 2002. *Corpora in Applied Linguistics.* Cambridge, UK and New York: Cambridge University Press.

Mickel, Stanley L. 1999. *Dictionary for Readers of Modern Chinese Prose: Your Guide to the 250 Key Grammatical Markers in Chinese.* New Haven: Far Eastern Publications, Yale U.

Okurowski, Mary Ellen and John Kovarik. 1998. The Chinese Treebank Project 100 Sentences Home Page: 汉森林 Chinese Forest -- The Chinese Treebank Project 小丛树 The 100 Sentence Corpus. URL: <umiacs.umd.edu/labs/CLIP/tocampsen.html>. (URL for the 'Segmentation Guidelines: <umiacs.umd.edu/labs/CLIP/forest.html>.)

Rodriguez, Maria Rosario Caballero. (n.d. (early 1990's?). Using a Concordancer in Literary Studies. (Online version at: <vlc.polyu.edu.hk/Concordance/ Review/programa.htm>.)

Stevens, Vance. 1995. Concordancing with Language Learners: Why? When? What?" *CAELL Journal* 6.2: 2-10. (Online version at: <www.ruf.rice.edu/~barlow/stevens.html>.)

Tao, Hongyin. 2000. Adverbs of Absolute Time and Assertiveness in Vernacular Chinese: A Corpus-Based Study. *Journal of the Chinese Language Teachers Association* 35.2: 53-73.

Teng, Shou-hsin (ed.). 1996. *Chinese Synonyms Usage Dictionary*. Boston: Cheng & Tsui Company.

Tribble, Chris, and Chris Jones. 1990. *Concordancing in the Classroom: A Resource Book for Teachers*. London: Longman.

Wang, Yu-Fang. 1999. The information sequences of adverbial clauses in Mandarin Chinese conversation. *Journal of Chinese Linguistics* 27.2: 45-89.

Wu, Dekai. 1998. A Position Statement on Chinese Segmentation. Paper presented at the Chinese Language Processing Workshop, University of Pennsylvania, Philadelphia, 30 June to 2 July 1998. (Online version at: <www.cs.ust.hk/~dekai/papers/segmentation.html>.)

Yu, Hsiao-jung. 1996. Consistent inconsistencies among the interrogatives in *Rulin Waishi*. *Journal of Chinese Linguistics* 24.2: 249-280.

APPENDIX

Concordancing Software for Chinese E-Texts Running Under English Windows 2000

Four programs for concordancing of Chinese e-texts are compared in this appendix, with all four running under English Windows 2000; they have not yet been tested under Windows XP. The four concordancing programs, listed below, are compared with respect to operating system, sorting capabilities, and some other important features to consider (total 14 items).

SOFTWARE	INFORMATION
Wenlin	*Wenlin Software for Learning Chinese* (versions 2.5 and 3.0) • http://www.wenlin.com • Developer: Tom Bishop • commercial program (US ~$150.00) • demo available – non-expiring, limited dictionary
ConcApp	*ConcApp Concordance Browser and Editor for Windows 2.0* (English/French/Chinese/Japanese) • http://vlc.polyu.edu.hk/pub/concapp/concapp.htm • Developer: Chris Greaves • freeware
MonoConc Pro	*MonoConc Pro 2.0* • http://www.athel.com • Developer: Michael Barlow • commercial program (US $85.00) • demo available – non-expiring, limited hits
Concordance	*Concordance 3.0* (pre-release version 2.9.9 was used here) • http://www.rjcw.freeserve.co.uk • Developer: Rob J.C. Watt • commercial program (US $89.00) • demo available – 30-day trial, fully-functional

1.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
Operating systems	DOS 3.0 or later, Windows 9x/NT 4.0 or later, and Mac System 7.0 or later	Version 1 for Windows 95 Version 2 for Windows 95x. (This program can also work under Windows 2000, but displays better under Windows 9x than under Windows 2000.)	Windows 3.1/ 9x/2000/Me (This program works best under Windows 9x. It can work under Windows 2000, also, but the source e-text window cannot decode and display Chinese characters.)	Windows 9x/NT/ 2000/Me (This program works best under Windows 2000 for Chinese, as Windows 2000 has better multilingual support than Windows 9x/Me.)

2.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
E-texts as corpus: Encoding systems	GBK/GBK, Big5/Big5+, HZ, UTF8, and UTF16 (i.e., Little-endian and Big-endian Unicode)[1]	GB/GBK, Big5/Big5+	GB/GBK, Big5/Big5+	GB/GBK, Big5/Big5+

[1] *Wenlin* describes HZ as 'disguised GB.' HZ-encoded e-texts have not been tested here with these four programs. Note that *Wenlin 3.0* supports Unicode 3.1 whereas *Wenlin 2.5* supports Unicode 2.1. The Arial Unicode MS font, for example, contains the approximately 40,000 alphabetical characters, CJK characters, and symbols that are defined in the Unicode 2.1 standard, and can, in lieu of a Chinese font, be the font selected for displaying Big5/Big5+ and GB/GBK in Chinese concordances in *MonoConc Pro 2.0* and *Concordance 3.0* running under English Windows 2000.

3.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
E-texts as corpus: File types	Text files (with encoding listed in item 2).	RTF, DOC, Text (The program can also re-open concordance databases saved in .mdb format, supported by MS Access database management program.)	Text files (with encoding listed in item 2). HTML files: HTML-tags can be suppressed in the Context window. POS-tags: Can be suppressed.	Text files (with encoding listed in item 2). HTML files: HTML-tags can be suppressed in the Context window. POS-tags: Can be suppressed.

4.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
E-texts as corpus: Number of source files as input	One or more. Wenlin cannot highlight to select more than one file in a subdirectory, but wildcards (*) can be used (e.g., *.gb, *.*).	One or more files can be opened one at a time. Wildcards (*) cannot be used.	Can highlight and select one or more files. Multiple files are concatenated into a single e-text.	Can highlight and select one or more files. Multiple files are concatenated into a single e-text.

5.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
E-texts as corpus: spaced e-texts required	No. Searches for Chinese characters (and English words) are handled as searches for strings of text, i.e., regular expression searches.	No. Searches for Chinese characters are best conducted as searches for 'any string', i.e., a regular expression search under English Windows.	No. For spaced e-texts, searches for Chinese characters can be treated as words, phrases, or strings. For non-spaced e-texts, searches for Chinese characters need to be conducted as searches for strings (regular expression searches).	No. For spaced e-texts, fast concordances (of words, phrases or strings), as well as full concordances, can be generated. For non-spaced e-texts, searches for Chinese characters need to be conducted as searches for strings, i.e., regular expression searches.

6.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
External encoder/ decoder required	No. Not needed for either display or encoding for conducting searches.	Yes. Needed for display of Chinese characters and encoding for conducting searches when the program is run in English Windows.	No Not needed for display if a Chinese font is selected.[2] CJK-input support built into Windows 2000 can be used for	No Not needed for display when a Chinese font is selected. Not needed for making full concordances.

[2] In *MonoConc Pro 2.0*. running under Windows 2000, the window displaying the source e-text cannot decode the Chinese, with or without an external decoder. There is no font selection for that window, unfortunately. In Windows 98, Chinese characters can be displayed in all windows.

6.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
			inputting characters for word and string searches, or an external decoder.	An external decoder, or CJK-input support built into Windows 2000, can be used for inputting characters for word and string searches.[3]

7.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
KWIC display format **(Keyword in Context)**	No. Wenlin is not a dedicated concordancing program.[4]	Yes. Highlight color for the keyword. In addition, the keyword is highlighted and displayed in the source e-text in a separate window.	Yes. Highlight color for the keyword. In addition, the keyword is highlighted and displayed in the source e-text in a separate window.	Yes. Separate, user-adjustable column width for the keyword, centered in the column. In addition, the keyword is highlighted and displayed in the source e-text in a separate window.

[3] The program has font selection options to be applied throughout the program for correct display of Chinese characters in all windows. In addition, font selection options are available for some specific windows to enable selection of different font face, font size, font color, etc. In the case of the View window displaying the source e-text, font selection is limited to fixed-width fonts.

[4] Without the tokens of the keyword in KWIC display format, the tokens may occasionally be at, or near, the beginning or end of the context displayed, so that one might not be able to view the collocates.

8.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
Single vs. multiple word searches **Use of wildcards**	Single word and phrase searches only (i.e., searches for strings). Does not support use of wildcards.	Single word and phrase searches via searches for strings. Does not support use of wild cards.	Single and multiple word/phrase searches via searches for strings with regular expression searches. Supports use of wild cards (*, ?, etc.) for full, regular expression searches for strings and substrings, for tagged searches, etc.	Single word and phrase searches via searches for strings using regular expression searches of spaced and non-spaced e-texts. Single and multiple word/ phrase searches using pick lists with spaced e-texts. Supports use of wild cards (*, ?) for advanced searches of strings and substrings, etc.

9.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
Full concordances of spaced e-texts[5]	No. Can only search for a specific word	No. Can only search for a specific word	No (not really) It is designed primarily for conducting	Yes.

[5] As of January 2002, none of these programs can make full concordances using non-spaced Chinese e-texts.

[6] As noted in item 8, *MonoConc Pro 2.0* can search using wildcards. As a result, a simple search of the asterisk (*) on a character-spaced e-text, for example, would yield a full concordance of every Chinese character in the corpus. The search results are accomplished via mechanically selecting the first word in the source e-text, then the next word, and so on, until it has searched to the end of the e-text. At the same time, all keywords are displayed together in the Context window in KWIC format.

9.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
	or phrase (i.e., a search for a string).	phrase, or string.	searches of words, phrases, and strings.[6]	

10.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
Line numbering of search results	No. However, total number of occurrences, or hits, is given. Filename and line numbering based on the lines of the source e-text optionally included in the search results.	Yes. This is an optional selection. In addition, total number of hits, is given.	No. However, total number of matches, or hits, is given.	Yes. In addition, total number of tokens, or hits, is given for each searched item. Line numbering is given as part of the concordance based on the lines of the source e-text.[7]

11.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
Sorting Capabilities	Cannot sort the searched results.	Can sort to the left or to the right of the keyword. However, sorting can only be done on up-	Can conduct sorting by keywords and by contexts to the left and right of the keyword, as well as other	Can conduct numerous sorting sequences for the keywords, including sorting by order of occurrence, fre-

Note also that the program cannot sort by type for frequency of tokens to type, whereas other concordancers, such as *Concord* (in *WordSmith Tools 3.0*) and *Concordance 3.0*, for example, can.

[7] 'References' can be added to source e-texts for subdividing the text into chapters, speaker turns, etc., and such information can be sorted in the Context window, where keywords are shown in KWIC display format.

11.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
		sorted concordances; the results of a concordance search that have been sorted cannot be re-sorted by the program.	advanced sorting routines, but the program cannot sort by frequency of occurrence.	quency, word-ending, by 'alphabetical order,' etc.[8] Multiple sorting capabilities are also available for the Context window displaying KWIC format.

12.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
Collocations	No	Yes. Left or right collocates of the keyword are obtained from the output of sorting of left or right collocates.	Yes. A collocation of the keyword can be made, with two collocates to the left and two to the right of the keyword.	Yes. A collocation of the keyword can be selected, with collocates to the left and to the right—four words to the left and four words to the right of the keyword.

[8] 'Alphabetical' sorting of 'strings' in *Concordance 3.0* is quite useful for sorting of concordances made from GB/GBK-encoded e-texts, because the sorting is based on the internal ordering of the encoding system. In GB (GB 2312-80), Chinese characters are ordered based on Pinyin romanization, including tones, for the 3,755 most frequently-occurring Chinese characters (at *hanzi* level 1), and the remaining 3,008 Chinese characters are ordered by Radical/Stroke count (i.e., first by the 214 radicals (部首) and then by stroke count (at *hanzi* level 2). (See Steven J. Searle's webpage on 'A Brief History of Character Codes in North America, Europe, and East Asia' at <tronweb.supernova.co.jp/characcodehist.html>.) GBK (GB13000), the expanded GB encoding system containing some 21,000 Chinese characters, retains the two-tier ordering scheme. The ordering of Chinese characters in Big5 (containing some 13,000 Chinese characters), however, is less systematic and hence less useful. Unicode e-texts are not yet readable in the program, but in a future version that can handle Unicode/UTF8-encoded e-texts, alphabetical sorting of strings will result in ordering of Chinese characters based on Radical/Stroke count.

13.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
Save as file types[9]	Text files in Big5+; GBK; UTF7, UTF8, and Unicode (Little- and big-endian Unicode). Search results can also be copied and pasted to a Unicode-based programs, (e.g., MS Word 97, 2000).	RTF, DOC, and Text files encoded in Big5 or GB. Concordances saved as concordance databases (in MS Access' .mdb format) can be re-opened for further querying later.	Text files in Big5 or GB encoding.	Text file, HTML, and Concordance files encoded in Big5 or GB. Files saved as concordances, with file extension '.concordance', can be re-opened later in the program for further study, sorting, etc.

14.	WENLIN	CONCAPP	MONOCONC	CONCORDANCE
Web Concordances	No.	No.	No.	Yes. Web concordances can be built from fast and full concordances generated by the program, from which online searches can be conducted.

[9] Concordances saved as text files, as well as those in RTF and DOC format, cannot be re-opened in these concordancers for further manipulation. That is, these files would simply be treated as source e-texts by the programs.