Journal of Technology and Chinese Language Teaching Volume 7 Number 1, June 2016
http://www.tclt.us/journal/2016v7n1/li.pdf pp. 56-77

Collocation Analysis Tools for Chinese Collocation Studies (可用于汉语搭配研究的搭配分析工具)

Li, Shouji
(李守纪)
Massey University
(梅西大学)
s.li.1@massey.ac.nz

Guo, Shulun
(郭曙纶)
Shanghai Jiao Tong University
(上海交通大学)
gshulun@163.com

Abstract: The importance of the teaching and learning of collocation has been widely acknowledged among researchers in the field of second language acquisition and teaching. In recent years, a number of lexical tools that can be used for collocation analysis such as WordSmith, AntConc, PowerConc, BNCweb, and CQPweb have become available for researchers. Despite their availability, the understanding and mastery of these tools and related techniques remain challenging for many researchers who are technically less competent. This is particularly the case in the studies of Chinese collocation within the field of Teaching Chinese as a Foreign Language (TCFL). Some of the abovementioned tools cannot be used for processing and analyzing Chinese texts, and even if they could, extra steps would be needed before they could be employed. This study aims to provide an overview of the collocation tools and corpora for Chinese language learning, with a focus on tools that can be used for Chinese language. Using the corpus analysis toolkit AntConc as an example, the study explains the procedures involved in the analysis of Chinese collocations, the configurations of the toolkit, and the statistical measures that are relevant to Chinese text analysis. It is hoped that this introduction provides some useful information for researchers who are interested in using collocation analytic tools in their studies.

摘要：搭配教学的重要性近年来已经得到研究第二语言习得和第二语言教学界的普遍认可。相应地，可以用于搭配研究的一些词汇学工具例如 WordSmith, AntConc, PowerConc, BNCweb, CQPweb 等也已经开发出来并为诸多学者所使用。然而，对于很多缺乏这方面技能的研究者来说，理解并掌握如何利用这些工具进行搭配研究却成为一个挑战。这种情况在汉语搭配研究当中更为突出，原因是上述很多工具最初是针对分析英语文本而设计的，无法直接用于汉语搭配的分析，或者是有些工具虽然可以用于汉语文本分析，但使用前需要一些额外的步骤对汉语文本进行处理。本文首先概述了可以用于汉语搭配研究的词汇工具和语料库的基本情况，然后以语料库分析工具 AntConc 为例，详细说明了使用它对汉语语料中的搭配进行提取和分析时的操作步骤、软件设置、跨距和统计测量值的设定等，以期对那些有兴趣使用这些搭配分析工具进行汉语搭配研究的学者提供一些有用的信息。

Keywords: Collocation studies, Collocation analysis tools, Chinese corpus, TCFL

关键词：搭配研究，搭配分析工具，汉语语料库，对外汉语教学

1. Introduction

In recent years, the importance of the teaching and learning of collocation has been widely acknowledged among researchers in the field of second language acquisition and teaching. As a subset of formulaic sequence, collocation is regarded by some scholars as one of the most important aspects of language teaching and also one of the most significant challenges faced by second language learners. For example, Palmer (1981) regards collocation as a crucial key to language learning. In his opinion, "language are learned collocation by collocation rather than word by word" (p. 21). Likewise, Lewis (2000) believes that "collocation is the most powerful force in the creation and comprehension of all naturally-occurring text" (p. 45). Nesselhauf (2003) stresses that collocations are especially important for learners who want to achieve a high level of proficiency in their target language because they help to enhance both accuracy and fluency.

Despite the realization of the importance of collocation in language acquisition and learning, the definition and identification of collocations remain a challenge for many researchers. Over the years, there have been various taxonomies proposed by researchers with regards to how to define, identify, and distinguish collocations based on multiple criteria. These views, however, can be largely summarized as belonging to two major groups: frequency-based views and phraseological views. The frequency-based view, represented by Sinclair (1991), defines collocation as "the occurrence of two or more words within a short space of each other in a text" (p. 170). This definition is based on the probability of co-occurrence of two or more lexical items and is supported by statistical data obtained from the analysis of large language corpora. The second approach to defining collocation, namely the phraseological view, is based on the analysis of syntactic structure and semantic motivation of co-occurred lexical combinations using certain criteria. Each of these two approaches has its own pros and cons. The data-driven approach runs the risk of identifying some recurring lexical clusters (such as *and the*, *of a*) that have little psycholinguistic validity to native speakers, and the phraseological approach lacks the statistical evidence of the frequency of their actual use in language communication (Henriksen, 2013, p. 31). To overcome these disadvantages, researchers usually apply both approaches in their studies on collocations; first identifying high-frequent lexical bundles in large scale language corpus using a set of statistical measures and subsequently excluding those combinations that do not meet the collocational criteria according to the analysis of their syntactic structures and semantic motivations. It is, therefore, safe to say that the use of collocation analytical tools and language corpus has become increasingly indispensable for collocation studies.

2. **An overview of collocation analysis tools**

With the development of technology in the last few decades, a number of corpus analysis tools that can be used for collocation analysis have become available to researchers. These tools include WordSmith, MonoConc, AntConc, PowerConc, BNCweb, JustTheWord, COCA, TANGO, the Gutenberg Collocation Tool, Wmatrix, SketchEngine, Phrase in English and CQPweb. According to Hardie (2012), these concordance tools can be classified as four generations according to their power, flexibility, and usability. The first-generation tools are mainframe-based software such as the CLOC (by Reed, 1978) concordancer used at the University of Birmingham, and the second-generation are mainly PC software such as the Kaye concordancer (developed by Kaye, 1990), the Longman Mini Concordancer (by Chandler, 1989) and Micro-OCP (by Hockey, 1988). The third generation tools, represented by WordSmith (developed by Scott, 1996), MonoConc (by Barlow, 2000) and AntConc (by Anthony, 2005), were designed partially to meet the need for tools that can be used by the majority of linguists who are less technically competent and lack programming skills (Hardie, 2012, p. 382). The second- and third-generation corpus tools are usually desktop computer-based programmes, and are usually under a Windows, Macintosh or Linux operating system.

These corpus tools, however, as Hardie (2012) argues, have some limiting factors in terms of power and usability. The term 'power' refers to the capability of achieving high query speed on very large corpus datasets. The second- and third generation corpus tools often lack the power to handle large-scale corpus data that researchers need to achieve their goals due to the processing ability of computer hardware. Usability refers to user-friendliness, which has been a critical factor that constrains non-technical linguists from using these analysis tools. Hardie (2012) points out that a general relationship between power and user-friendliness can be summarized as that the more powerful a corpus analysis is, the less user-friendliness it has. For example, although WordSmith is a usability-oriented tool, one of its most powerful features is actually the indexing process, which is often less accessible for most non-technically-savvy corpus analysts. To address these issues, the fourth generation corpus tools use a different approach to improving both their power and usability by using a client/server model. According to Hardie (2012), CQPweb is one of the best web-based fourth generation corpus analysis tools, because it can enable non-technical corpus linguists to carry out corpus-based data analysis much like browsing web pages. It not only offers all the functionalities that most second- and third generation tools offer such as word list, concordancing, collocation, and keyword analysis, but also allows users to search for metadata of texts such as pre-defined genres, speakers' language proficiency and gender information. Furthermore, it provides corpus specialists with more advanced search options so that they can make more sophisticated queries. In terms of its language dependency, CQPweb has two additional features that are more attractive to corpus linguists whose research objects are languages other than English. According to Hardie (2012), CQPweb is "both language-independent and writing-system-independent" and "corpora in the Arabic, Bengali, Chinese, Cyrillic, Devanagari and Latin writing systems have been analyzed using CQPweb" (p. 398). Other highly regarded fourth generation tools include Wmatrix (developed by Rayson, 2008), SketchEngine (by Kilgarriff et al., 2010), and the corpus.byu.edu system (by Davies 2005, 2009, 2010). All the aforementioned analysis

tools can be used for collocation analysis based on English corpora, with a few specifically designed for processing and analyzing collocations in English, such as JustTheWord, TANGO, and the Gutenberg Collocation Tool.

With regards to whether they handle Chinese text analysis, only WordSmith, MonoConc, AntConc, SketchEngine, and CQPweb have the capability to analyze collocations in Chinese corpora. Some later versions of WordSmith, MonoConc, AntConc are capable of processing and analysing Chinese texts if they are encoded in Unicode and tokenized/segmented in advance. The fourth generation tools Sketch Engine and CQPweb can also be used for collocation analysis based on corpora that have already been provided on their servers. For example, Sketch Engine provides four Chinese corpora: Chinese GigaWord 2 Corpus (mainland, simplified, 250,124,230 tokens); Chinese Giga Word 2 Corpus (Taiwan, Traditional, 455,526,209 tokens), ChineseTaiwanWac (Traditional, 349,198,060 tokens) and ChineseTaiwanWac (Universal Sketch Grammar, 465,102,710 tokens). Users can also create their own corpora by uploading files from their own computer or download corpus datasets from FTP sites. Since CQPweb is an open source corpus query system, users can create their own corpora to use. For example, Xu (2014) created a BFSU CQPweb, which has 35 corpora in seven languages including Chinese and can be accessed at http://111.200.194.212/cqp/. Thus far, BFSU CQPweb has four original Chinese corpora: Lancaster Corpus of Mandarin Chinese version 1 (LCMCv1, Brown family, 1991), Lancaster Corpus of Mandarin Chinese version 2 (LCMCv2), TORCH2009 (Texts of Recent Chinese, Brown family, 2009, 2013 summer edition), and the UCLA Corpus of Written Chinese (2nd edition).

When choosing the most suitable tools for their collocation studies, researchers need to consider their research goals, available datasets or corpora, as well as the limits of available corpus analysis tools. According to Xu & Jia (2013), the third- and fourth generation of corpus analysis tools will co-exist for a period of time due to their advantages and limitations. They suggest that, in consideration of the researchers' actual needs, the third generation tools are probably more suitable for carrying out individual corpus-based studies. As a third generation tool, WordSmith has a complex interface and does not support regular expressions; in comparison, AntConc is easier to use and supports regular expressions, but it has less functionality and computational efficiency. It also tends to freeze or quit unexpectedly when processing larger corpus data. The fourth generation tools, as discussed above, are Internet-based network applications. Such tools are based on data and index technology, having faster retrieval response time and offering better user experience. When supported by the necessary hardware such as powerful servers, they are more capable of handling large corpus such as BNC. However, as the flexibility of these tools is not sufficient, users usually have difficulties with processing and analyzing corpus data stored on local desktop computers rather than a web server.Due to the limits of the index format and the scale of the data, their retrieve grammar is relatively simple, and they do not support complex searches. Considering all the factors, the authors choose AntConc to demonstrate the process and procedure involved in conducting Chinese collocation studies.

It is worth noting that the list of collocation analysis tools discussed here is not exhaustive. There are many other tools that might be more suitable for an individual

researcher's purposes. The aim here is to provide some useful information for researchers who are less technically competent but would like to gain a basic understand of such tools.

3. An overview of Chinese corpora and interlanguage corpora

3.1 List of corpora

In recent years, there have been major advancements in natural language processing. This enables more language corpora to be established for linguistic studies. To the best of the authors' knowledge, there are a few large scale Chinese corpora and Chinese learner corpora (or interlanguage corpora) freely available for language researchers. Below is a list of corpora that can be used for Chinese collocation studies:

(1) The BCC Corpus

The BCC Corpus (Beijing Language and Culture University Corpus Center) is a contemporary Chinese corpus developed by the Institute of Big Data and Education Technology of Beijing Language and Culture University. It has a 15 billion character collection of text samples of present-day written language from various sources including microblogging, science and technology, literature, and the press. The BCC Corpus provides an online concordancer which can be accessed at http://202.112.195.249/bcc/. Users can use search query to extract the use of words, grammatical markers, and other unique units such as separable words (离合词). It offers a statistical function, which not only allows users to search collocations using some formulaic expressions, but also provides statistical data of frequencies of different collocates. This is particularly useful if researchers want to find the frequency of certain collocates used in a specific text genre. For example, if one wants to know how the word 良好 collocates with other nouns in news texts, he or she can use the formulaic expression 良好* n to search. The following screenshot shows the search result of all the possible collocates of 良好。

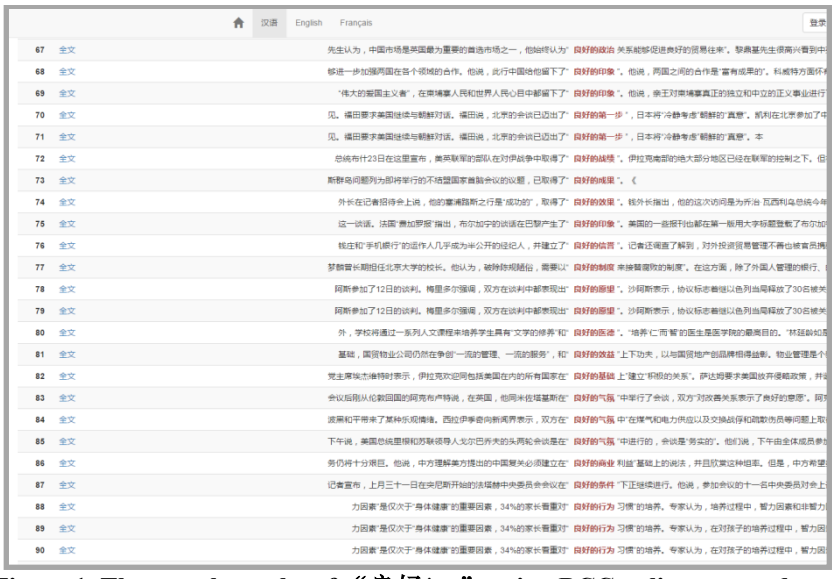


Figure 1. The search results of “良好* n” using BCC online concordance

Figure 2. The statistical data of nouns that in “良好* n”

The list is useful as it gives the frequency of each noun that can enter into the collocational phrase “良好的 n”. The list can also be downloaded for further analysis. However, one point that needs to be noted here is that some of these figures have to be manually checked for accuracy. For example, the list shows that 良好的社会 is the most frequent phrase used in the news texts with a frequency of 2581. But after a careful observation of the contexts 良好的社会, one will find that 社会 is actually not the collocate of 良好; it is only part of the nominal modifier. See the following examples:

(a)…在进行结构改革的同时还必须建立良好的社会保障。
(b) …没有良好的政府信用，就绝不可能建立起良好的社会信用。

(c) …提高服务质量，赢得了良好的社会信誉。

It is clear that in (a), the collocate of 良好 is 保障，in (b) is 信用 and (c) is 信誉。

(2) The PKU-CCL Corpus

The PKU-CCL Corpus is an online Chinese language corpus with 477 million characters, a collection of Chinese written texts of different genres. The corpus can be accessed at http://ccl.pku.edu.cn:8080/ccl_corpus/. This online query system supports word search, formulaic expression queries as well as queries of some unique patterns in Chinese, such as 高高兴兴(usually summarized as AABB pattern of duplicated adjectives), therefore it can also be used for Chinese collocation studies. However, it does not provide statistical functions similar to the BCC corpus provides.

(3) General Contemporary Chinese Corpus

This corpus is also an online Chinese language corpus with a token size of 19455328 and can be accessed at http://www.cncorpus.org/. It is sponsored by the State Language Affairs Committee of the Ministry of Education of China. Although this corpus only provides users with word or words online queries, it offers some very useful and free corpus analysis resources for users to download. For example, it offers the segmentation and POS tagging tool CorpusWordParser.ext and the word frequency tool CorpusWordFrequencyApp.ext, which are very useful in collocation studies. In Section 4, the authors describe how CorpusWordParser.exe can be used to prepare Chinese text for further collocation analysis using AntConc.

(4) The LIVAC (Linguistic Variations in Chinese Speech Communities)

The LIVAC corpus is a synchronous corpus developed by Hong Kong City University. It aims to offer a system that can be used to store the data and to analyze the linguistic development of printed Chinese texts in difference Chinese communities. The corpus also provides an online query system that can be accessed at http://www.livac.org/search.php.

(5) Academia Sinica Balanced Corpus of Modern Chinese

Sinica Corpus is a balanced corpus developed by the Academic Sinica in Taiwan with 10 million words (character token size 17,554,089). All the texts in this corpus are collected from different genres and categorized according to five criteria: genre, style, mode, topic, and source. Every text is segmented and every word is POS tagged. It also offers a web-interface which can be accessed at http://www.sinica.edu.tw/SinicaCorpus/. The online query system is designed for statistical comparison according to users' specification of topics, genres, etc. It is worth noting that this the characters in this corpus is traditional Chinese characters rather than simplified ones.

(6) ToRCH2009: Texts of Recent Chinese

The Corpus is also known as the 2009 Brown family Chinese corpus ToRCH 2009. It is the acronym of 'Texts of Recent CHinese'. The name ToRCH was proposed by Xu Jiajin and the corpus was released in summer 2013. ToRCH 2009 contains texts of 15 types (Press: Reportage, Press: Editorial, Press: Reviews, Religion, Skill and hobbies, Popular lore, Belles-lettres, Miscellaneous: Government & house organs, Learned, Fiction: General, Fiction: Mystery, Fiction: Science, Fiction: Adventure, Fiction: Romance, and Humour). The corpus size in tokenised words is 1,066,347 (1,670,356 Chinese characters) and one of the editions of ToRCH 2009 was already tokenised/segmented using ICTCLAS2012. It can be downloaded at http://www.bfsu-corpus.org/channels/corpus as a corpus stored at a local device such as desktop computers; therefore, it is a good dataset for researchers who want to perform collocations studies according to their own research goals.

(7) The HSK Dynamic Composition Corpus

The HSK Corpus is a collection of 11,600 essays (approximately 4.3 million Chinese characters) written by learners of Chinese for the HSK test. It was developed by the Research Center for Studies of Chinese as a Second Language at Beijing Language and Culture University. The HSK Corpus not only annotates the error information of characters, words, sentences, and text features such as cohesive devices, but also provides some possible corrections to these errors marked with tags. Some important text attributes such as nationality, gender, and age of CFL learners who took the test are also provided for researchers to consider when conducting studies. Statistical data pertaining to the use of characters, vocabulary, sentences, and discourse contained in compositions are also available for researchers. The corpus has an online query system which can be accessed at http://202.112.195.192:8060. Researchers can extract erroneous use of characters, words, collocations, and cohesive devices by using this online query interface.

(8) Chinese Learners Corpus

Developed by the Advanced Center for the Study of Learning Science of National Taiwan Normal University, the Chinese Learners Corpus collects written text samples from learners of Chinese at different levels with 40 L1 backgrounds and contains 3 million characters. It separates the texts written for assignments and for exams. This is particularly useful for studies on contributing factors of second language writing quality and writing strategies. Information for this corpus can be found at http://advancedcenter.top.ntnu.edu.tw/achievement5_1.html.

Due to the limit of space, the authors can only list these corpora which are currently available. This list of corpora is by no means exhaustive. For more information about Chinese corpora and Chinese interlanguage corpora, please refer to Wu & Li (2009).

3.2 The pedagogical value of the corpora

Among these corpora, the HSK Dynamic Composition Corpus, the BCC corpus, and the Chinese Learner Corpus probably have much more pedagogical value than the rest of the corpora listed above, although the last one contains only Traditional Chinese characters, which may not be suitable for learners who learn Simplified ones. In a TCFL classroom where the access to such corpora is available, teachers can first encourage learners to search for the use of some confusing words in the HSK Dynamic Composition Corpus, or teachers can show learners a few erroneous instances found in the HSK corpus. The learners can then use the BCC corpus as a reference corpus to search forthese words' high-frequency collocates. As described in Section (1), BCC corpus has the statistical function of listing all the high-frequency collocates for users. By studying the erroneous use of the words and their high-frequency collocates, learners are given opportunities to observe and interpret the collocational patterns of these words. On the basis of such language experience, learners are able to develop their competence in identifying errors, classifying various types of collocates, and generalize to a more abstract pattern. Such an approach to teaching vocabulary has been regarded by some scholars as beneficial and valuable because it allows learners "to observe *what* is typically said in given circumstances, and *how* it is typically said, and to relate the two" (Sinclair, 2004, p. 18). These three corpora also serve as very valuable datasets for TCFL teachers to conduct classroom-based studies. For example, Li (2016) investigates the collocational errors in compositions written for the HSK by American learners of Chinese on uses the HSK Dynamic Composition Corpus and the BCC corpus, and provides some useful strategies to improve the outcomes of the teaching of vocabulary. Due to the design principles and aims of data collection, other corpora listed in this study are probably more suitable for corpus linguistics studies with more specific research goals.

4. A demonstration of the use of AntConc for Chinese collocation study

As discussed in Section 2, most of the corpus analysis tools currently available are originally designed for English text analysis. However, some of them such as AntConc can handle Chinese text analysis if the text is prepared carefully in advance. In this section, the authors use AntConc as an example to demonstrate how this freeware concordance program can be used for Chinese collocation studies.

AntConc is a corpus analysis toolkit that can be used as a concordancer to extract multiple examples of words or phrases and their common collocates from a corpus for analysis. It is developed by Prof. Laurence Anthony from Waseda University in Japan and can be downloaded at http://www.laurenceanthony.net/software/antconc/. More information about this software can also be found at this website.

To use it for Chinese collocation studies, the encoding of Chinese characters needs to be UTF-8. This can be set up under the "characters encoding" category in the "global settings" tab of AntConc.

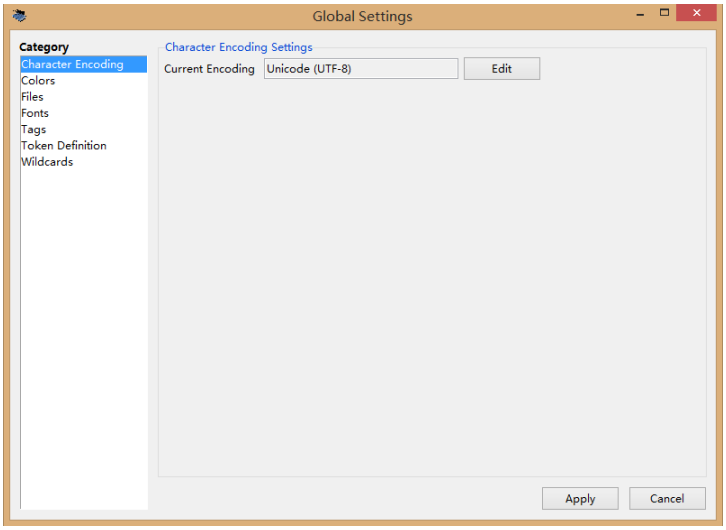


Figure 3. The encoding settings of AntConc

The next step is to set the statistical measures such as MI or T-Score. MI (mutual information) value and T-Score are two most frequently used statistical measures in collocation analysis. They are used to decide whether two co-occurred lexical items are by chance or whether their association is significant and has psychological validity. The Mutual Information value can be defined as the extent to which observed frequency of words occur in collocation differs from what we would expect. The T-Score can be defined as a measure that calculates the absolute frequency of co-occurrences of words. In his discussion of these measures, Stubbs (1995) points out that the issue of the MI value is that low-frequency collocates are more prominent in the MI based lists, while T-Score puts more emphasis on the number of joint frequencies. Consequently, more function words are likely to be included in a collocation list based on a T-Score than on MI, whereas lexical collocates that infrequently co-occur with the searched word are more likely to be found in a collocation list based on MI value(pp. 9-12). Most current corpus analysis tools provide either MI or T-Score, or both. Users are able to identify the collocates according to these two values if the MI value $\geqslant$3 and T-Score $\geqslant$2). It is worth mentioning that, according to Stubbs (1995), although the statistics can be generated mechanically by software or tools, users need to be aware of the differences of these measures and make their decisions accordingly.

According to Bai & Zheng (2004), to identify potential collocations in Chinese texts, the two measures needs to be set as MI $\geqslant$3, T$\geqslant$2.33, so co-occurrences with higher collocational strength can be highlighted. This can be set under the category of "collocates" in the "Tool Preferences" tab. Under the heading "Other Options" next to "Selected Collocate Measure", there are two options available: MI and T-Score. Choose one of them and click on "Apply" to complete the setting up process.

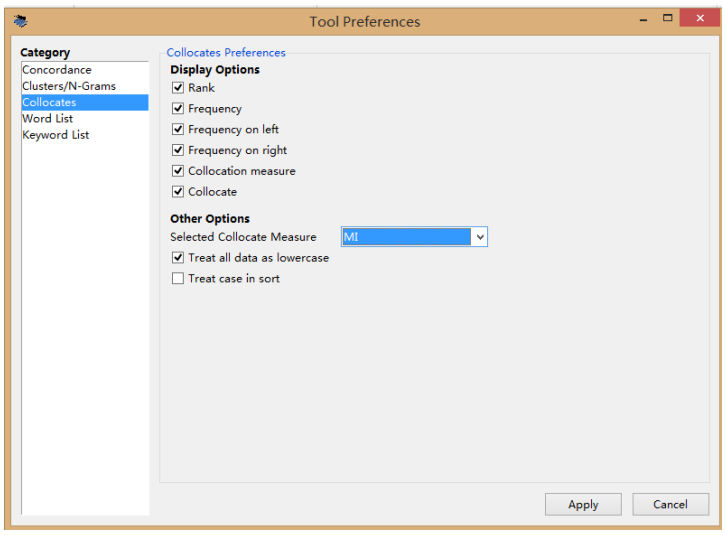


Figure 4. Setting the statistical measures

Once the set up is done, run AntConc and click on "File" on the menu bar. Choose "Open Dir…" , which will lead to the corpus data directory. See the following Figure 5.

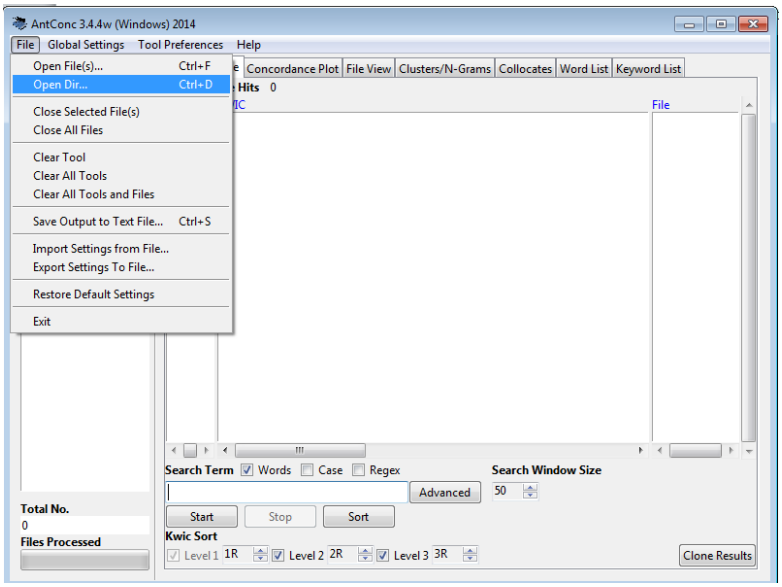


Figure 5. Importing the UTF-8 plain text files

Take the ToRCH 2009 corpus as an example. There are two sets of data available, one is ToRCH2009_ANSI—20140720, the other ToRCH2009_UTF-8—20140720. To search for collocates of a certain lexical item, just navigate to this folder that contains ToRCH2009_UTF-8—20140720 and click on "OK". The plain text files in UTF-8 encoding will be imported to the toolkit. If the plain text files in ANSI encoding are imported, the result will be "No Hit" or just a blank page after the files are processed. Once the right texts are loaded, the following settings need to done before performing the collocation query:

(1) Click on the "collocates" tab, and the relevant collocation settings will appear.
(2) Input the word to be investigated into the search box under the "search term" heading.
(3) Set the span of the words to the left and right of the searched term.

The default setting is 5L and 5R, which means 5 words to the left and 5 words to the right of the searched word. However, since the span is a critical concept of the statistical-based approach to studying collocation, the setting of span has much impact on the validity of the research findings. For example, Sun and Huang (1998) investigated the collocational distribution of noun (能力), verb (培训), and adjective (广泛) based on large scale Chinese corpus, and proposed that the best window spans for these three types of part of speech are: noun (-2, +1), verb (-3, +4), and adjective (-1, +2), here the figures in the parentheses mean the words to the left and to the right of the searched word, for example, (-2,+1) means two words to the left and one word to the right of the searched word. Setting these figures right may lead to more accurate results and retrieve rates. Bai (2004) and You (2005) also did similar research, and determined that it was a more suitable span for verbs and their collocates in Chinese text is (0,+5) as it covers most of the high-frequency collocations. It is, therefore, the researchers' decision to define the window spans to suit their research goals. For the purpose of demonstration, here the span is set as (0, +5) and search term (also known as *node* in some other studies) is a verb 提高。

(4) Decide how to sort the result. There are six methods to choose from: Freq, Freq(L), Freq(R), Stat, Word, Word End. Among them, the first four are more useful for the collocation analysis. "Freq" refers to the number of times the searched term and its potential collocates co-occur in the corpus, "Freq (R/L)" refers to the number of times the potential collocates occurs to the right or left of the searched term, and "Stat" refers to either MI or T-Score value which measures the collocation strength of the co-occurrence in the corpus. For this example, the result is sorted by Freq (R).

(5) The minimum collocate frequency can be set as the default setting "1".

After these settings are done, click on "Start", a warning message will pop up saying "AntConc needs to Jump to the word list tool to generate a word list". To calculate the collocation strength, it needs to know the frequencies of all the words in the corpus. Click on "ok" and the toolkit will start processing the data. A collocation report will be generated. See Figure 6.

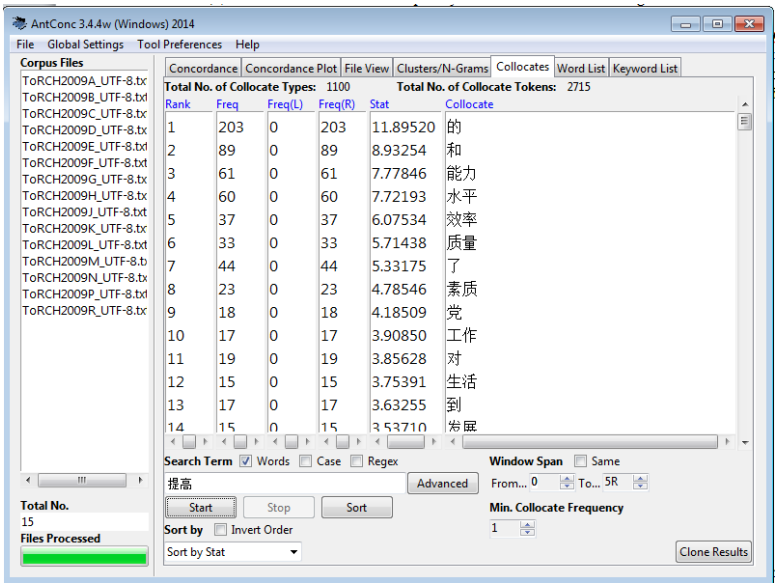


Figure 6. Collocation report of the verb 提高

It can be seen that the highest frequency and T-score value is 203 and 11.89520 respectively, and the word that collocates with 提高 is 的。However, this is just a high frequency recurring lexical bundle with little psycholinguistic validity, so words similar to this can be excluded from the high frequency list. We can see that the word 能力 has a high frequency at 61 and a high T-score at 7.77846. To view it in context, we can just click on the word 能力 and produce a concordance set of result. This step is essential as it will allow researchers to manually check if the co-occurrence has psycholinguistic validity and whether the word is a real collocate of the searched term. Figure 7 shows the concordance examples of the word 能力：

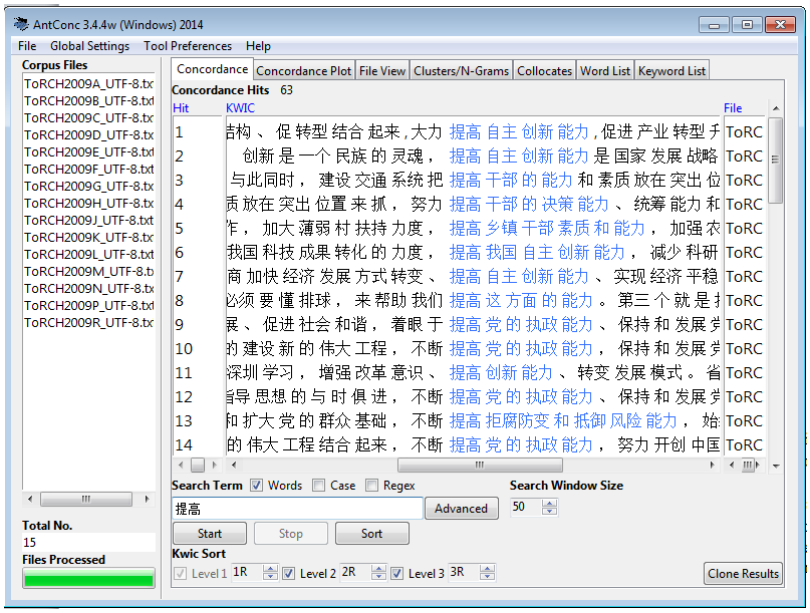


Figure 7. The concordance examples of 提高

One may also find that the word 党 does not look like a real collocate of 提高, although it has a high frequency at 18 and high T-Score at 4.18509. By checking its concordance result, we can see that the word 党 is actually part of the modifier for the nominal head words such as 水平, 能力 etc., see Figure 8. The word 党 therefore should be excluded from the high frequency list of 提高。

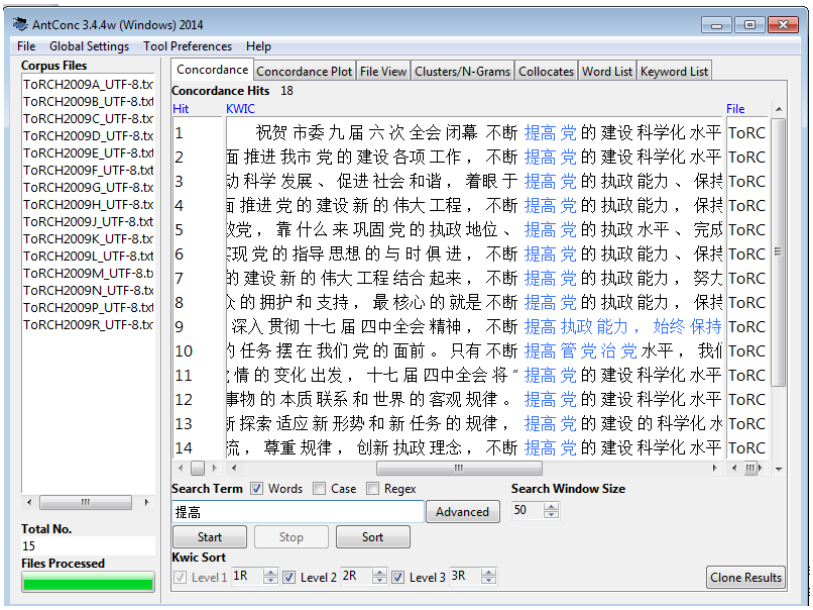


Figure 8. The concordance result of 党

To export the statistical data of the high frequency list of 提高 for further analysis, one can just click the File tab and choose "Save Output to Text File". The data will then be saved as a txt file.

Due to the relatively smaller size of the ToRCH 2009, the high frequency list of collocates is rather limited in terms of the insight it may provide of the features of a particular lexical item. For example, if we search for the high-frequency collocates of 丰富 from ToRCH 2009 and set the two measures as MI ⩾3, T⩾2.33, we only get the following list:

Table 1. High frequency collocates of 丰富 in ToRCH 2009

Collocate	MI	Frequency	Collocate	T-Score	Frequency
营养	8.88	6	营养	2.446	6
经验	8.38	6	经验	2.439	6
			内容	2.434	6
资源	6.76	5			
维生素	8.38	3			
资料	6.833	2			

To address this limitation, we may need to resort to larger scale corpus such as

BCC corpus mentioned in Section 3. However, due to the fact that BCC does not offer downloadable corpus raw data, it is impossible to use AntConc to perform collocation analysis. Even if it did, the sheer size of the corpus would impose a great difficulty for corpus analysis toolkits such as AntConc to process simply because of its limited power. One way to resolve this issue is to use BCC's online query system to generate concordance result of a particular lexical item, then use AntConc to perform further collocation analysis. Still using 丰富 as an example, we can search for the word on its online query page at http://bcc.blcu.edu.cn/. To limit the number of retrieved concordance results, it would be much more viable to choose one of BCC's sub-corpus such as literature corpus as the dataset (the system returned 31386 instances of 丰富), rather than its comprehensive corpus (returned 203210 instances of 丰富). The online query system allows registered members to download 10000 instances of the total results, which should be large enough for researchers to perform analysis and generate some valid high-frequency lists.

Once the concordance result of 丰富 is downloaded and saved as plain text file, the next step is to remove the <u> and </u> marks surrounding the word 丰富. The file is then parsed using ICTCLAS (Institute of Computing Technology, Chinese lexical Analysis System) or the POS tagging tool CorpusWordParser.ext downloaded from the General Contemporary Chinese Corpus. Figure 9 shows the screenshot of the interface of CorpusWordParser.exe when parsing the downloaded concordance result of 丰富。

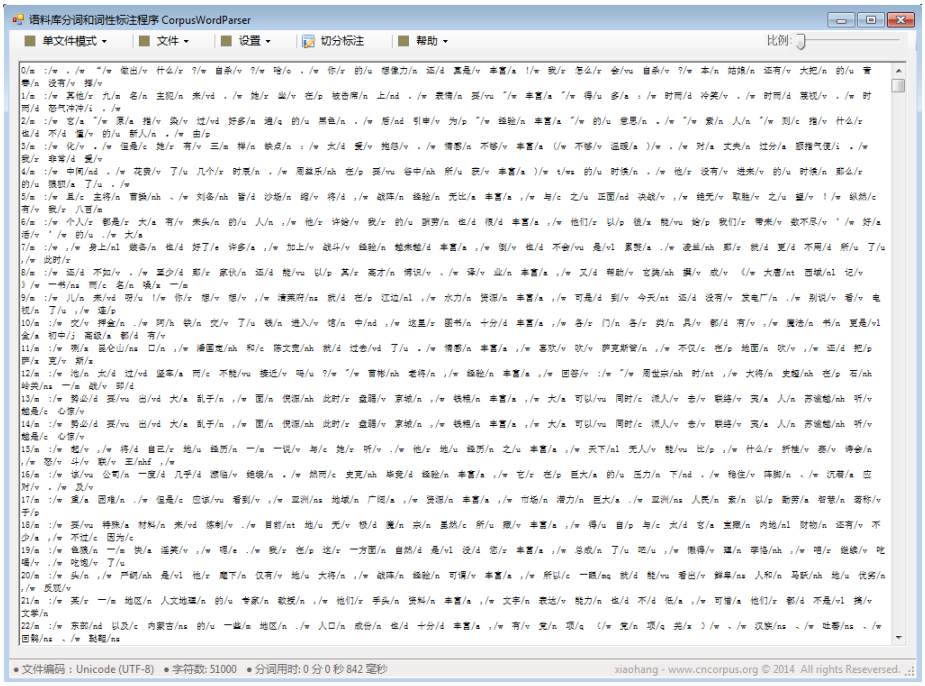


Figure 9. Using CorpusWordParser.exe to parse Chinese text

Once the text is complete, make sure to save it as a plain text file with UTF-8 encoding. See Figure 10.

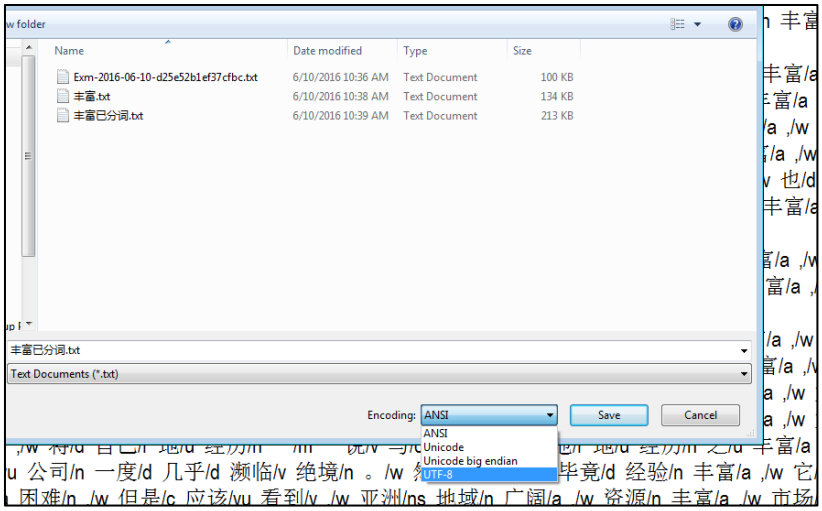
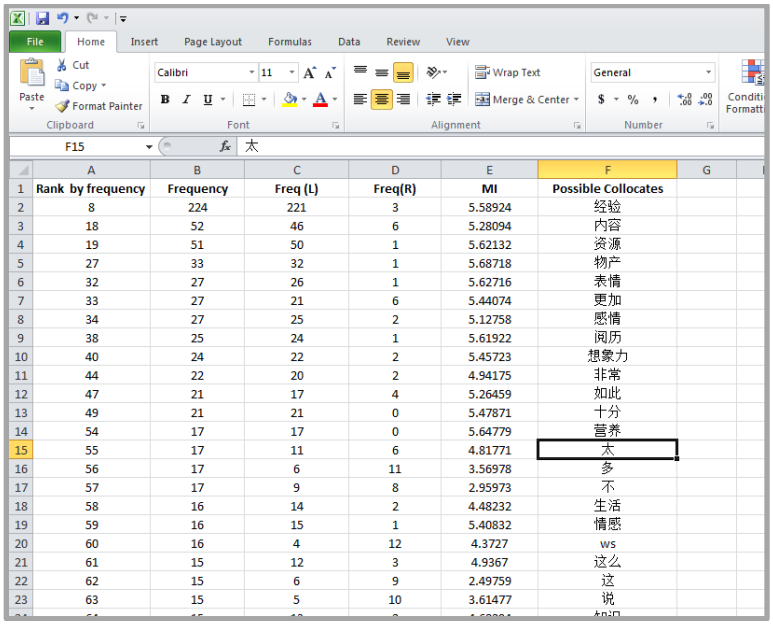


Figure 10. The parsed text needs to be saved as plain text file with UTF-8 encoding

Once the file is saved correctly, it can then be analyzed using AntCont. When setting the window span, the number of words next to the node word has to be doubled. AntConc treats the POS tag as a character. For example, if a researcher decides the window span for noun, verb, and adjective are noun (-2, +1), verb (-3, +4), and adjective (-1, +2) respectively, when he or she searches their collocates in parsed text, the window span should set as noun (-4, +2), verb (-6, +8), and adjective (-2, +4) respectively.

Once these are set, the text can be imported to AntConc and a collocation analysis performed. The analysis result can be exported as a plain text file. To conduct a further analysis, the result file can be imported into Microsoft Excel using the Data>From Text tool bar function. Figure 11 shows the process of analysis of the possible collocates of 丰富。



	A	B	C	D	E	F
1	Rank by frequency	Frequency	Freq (L)	Freq(R)	MI	Possible Collocates
2	8	224	221	3	5.58924	经验
3	18	52	46	6	5.28094	内容
4	19	51	50	1	5.62132	资源
5	27	33	32	1	5.68718	物产
6	32	27	26	1	5.62716	表情
7	33	27	21	6	5.44074	更加
8	34	27	25	2	5.12758	感情
9	38	25	24	1	5.61922	阅历
10	40	24	22	2	5.45723	想象力
11	44	22	20	2	4.94175	非常
12	47	21	17	4	5.26459	如此
13	49	21	21	0	5.47871	十分
14	54	17	17	0	5.64779	营养
15	55	17	11	6	4.81771	太
16	56	17	6	11	3.56978	多
17	57	17	9	8	2.95973	不
18	58	16	14	2	4.48232	生活
19	59	16	15	1	5.40832	情感
20	60	16	4	12	4.3727	ws
21	61	15	12	3	4.9367	这么
22	62	15	6	9	2.49759	这
23	63	15	5	10	3.61477	说

Figure 11. Using Microsoft Excel to further analyze possible collocates

As mentioned before, a manual check process is essential to ensure those lexical bundles which have little psycholinguistic validity are excluded. The final high-frequency collocate list of 丰富 (MI ⩾3) can be found in Table 2:

Table 2. High-frequency collocates of 丰富 in BCC literature corpus

Frequency	Freq (L)	Freq(R)	MI	Possible Collocates
224	221	3	5.58924	经验
52	46	6	5.28094	内容
51	50	1	5.62132	资源
33	32	1	5.68718	物产
27	26	1	5.62716	表情
27	21	6	5.44074	更加
27	25	2	5.12758	感情
25	24	1	5.61922	阅历
24	22	2	5.45723	想象力
22	20	2	4.94175	非常
21	17	4	5.26459	如此
21	21	0	5.47871	十分
17	17	0	5.64779	营养
17	11	6	4.81771	太
16	15	1	5.40832	情感
15	12	3	4.9367	这么
15	13	2	4.68294	知识
14	13	1	5.73025	极为
13	13	0	5.43069	收获
11	5	6	5.73025	肥沃
11	5	6	4.48924	思想
10	8	2	4.59275	那么
10	9	1	5.35174	极其
10	7	3	5.73025	学识
10	10	0	5.14529	不断
9	7	2	5.09282	物质
8	5	3	4.40832	经历
8	7	1	5.27082	材料
8	7	1	5.40832	内涵
7	5	2	5.36768	形式

Table 2 can be further analyzed according to the statistics. For example, one can clearly see that most of the nouns (经验, 内容, 资源, 物产, 表情, etc.) and adverbs (非常, 如此, 十分, 太, 这么, 那么, etc.) tend to appear to the left of the word 丰富, and the number of their Freq (L) are much higher than that of their Freq (R). Such information is useful for both TCFL teachers and learners of Chinese as this can give them a more

accurate and specific collocational pattern (Li, 2016).

If we compare Table 1 and Table 2, it is obvious that the size of a corpus is important for generating more valid research findings. Researchers need to consider carefully which corpus and corpus analysis toolkit best suit their research goals and make decisions accordingly.

A last note for the study is that due to the limit of space, this study does not provide introduction of the use of some fourth generation corpus analysis tools such as Sketch Engine (https://www.sketchengine.co.uk/) and BFSU CQPweb (http://111.200.194.212/cqp/), both of which are capable of performing Chinese collocations studies online. This may be offered in our future studies.

5. Implications for the teaching of vocabulary

In recent years a number of studies have been conducted on the use of corpus-based approach in L2 classrooms (Chan, 2002; Souza Hodne, 2009; Jafarpour, Hashemian, & Alipour, 2013). Some scholars also call this the data-driven learning approach. In this section, the authors use a few confusing words in Chinese to demonstrate how the collocation analysis can be helpful for learners not only to improve their collocational competence, reduce their erroneous use of collocations, but also to increase their awareness of collocations in the target language and therefore to produce more natural utterances.

In the HSK Dynamic Composition Corpus, some of the deviant use of 营造，造成，and 达成 are found:

(d) …男女分班制度由于缺少另一方而造成{CC 营造}了不平衡的生活圈子。
(e) …首先喜欢流行歌曲的人可以造成他们的同样的感情
(f) … 我认为流行歌曲可能会造成不好的文化，可流行歌曲的作用也不能忽视
(g) …可是对某些人类却造成{CC 达成}了很多负面的影响{CC 效果}。

It can be seen that these words are somewhat confusing to learners, as these words either share some constituents such as both 营造 and 造成 have a constituent 造，or they share a meaning that expresses “something is made to happen”.

In a TCFL classroom, teachers can ask learners to first search for these words’ high-frequency collocates using BCC corpus online concordancer, and then ask them to use the statistical function to list them on paper. Because main errors in the above sentences are the writers have used wrong nouns that collocates with verbs, teachers can ask learners to use regular expressions 营造*N，造成*N，and 达成*N to search for their instances in the corpus. Due to the space limit, only one of the screenshots of the statistical data is demonstrated here:

Figure 12. The high-frequency collocate list of 营造

Figure 12 shows that the high-frequency collocates include 环境，氛围，气氛，舆论，文化，空间，城市，优势，效果，景观， 市场，防护林等。 Among these words, most of them can be said to have an [ABSTRCT] semantic feature except 城市，景观，防护林。Teachers could then ask learners to check the modifiers of these nouns by click on the words listed in this statistical page. See the following screenshot:

Figure 13. Screenshot of the context of 营造*环境

A careful observation of the context can tell us that most of the abstract nouns are modified by words that have positive meanings such as 安全，公平，和谐，自然，良好，优美，健全 etc. Therefore, we can add one more semantic feature [POSTIVE] to the use of 营造*环境. The collocation pattern of 营造 therefore can be summarized as: Following such procedure, teachers can guide learners to work as groups to discuss and try to figure out this collocation pattern.

It is If TCFL teachers want the collocation pattern to be more accurate, they can use collocation analysis tools such as AntConc to conduct an analysis and generate a more reliable and comprehensive collocate list using the method we have demonstrated in the previous section.

Using the same procedure, teachers and learners should be able to work out the following collocation patters for the three words:

(1) 营造＋collocates [ABSTRACT/POSITIVE] except a few words that have concrete meaning such as 城市，景观，防护林 etc.

High-frequency collates: 环境，氛围，气氛，舆论，文化，空间，优势，效果，市场

(2) 造成 + collocates [NEGATIVE]

High-frequency collocates: 影响，死亡，损害，后果，浪费，危害，污染，伤害，伤亡，困难，问题，事故 etc.

(3) Two or more parties + 达成 + collocates [NEUTRAL]

High-frequency collocates: 协议，共识，一致，意向，意见， 妥协， 目的，交易，和解，默契，愿望，合作，效果, 契约 etc.

Using such lists, learners can work together to identify the reasons why sentences (d)-(g) are wrong, and then correct the sentences by themselves. Such process will help leaners to learn more productively about collocations and improve their collocation competence.

6. Conclusion

Corpus analysis tools and large scale corpora are becoming increasingly indispensable for studies of applied linguistics and second language acquisition. However, there are some challenges researchers face when employing these tools and resources. This is particularly the case for researchers who are less technically savvy. This study offers an overview of corpus analysis tools available for analyzing both English and Chinese text, and some of the most well-designed and well-compiled Chinese corpora that can be used for studies on Chinese collocation. The study then introduces the steps and procedures that are involved in using AntConc, a corpus analysis toolkit, to perform collocation analysis based on Chinese corpora such as ToRCH 2009 and BCC corpus. It is hoped that this demonstration may provide some useful information for those who are keen to employ similar tools in their own Chinese collocation studies.

References

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. In L. Anthony, S. Fujita, & Y. Harada (Eds.). *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning.* Paper presented at the IWLeL 2004: An interactive Workshop on Language e-Learning, Waseda

University, Tokyo, Japan, 10th December (pp. 7-13). Tokyo: Waseda University.
Bai, M., & Zheng, J. (2004). Study on ways of verb-verb collocation. *Computer Engineering and applications, 27*, 70-72. [白妙青, 郑家恒, (2004). 动词与动词搭配方法的研究. 计算机工程与应用, *27*,70-72.]
Barlow, M. (2000*). MonoConc Pro*. Houston, TX: Athelstan.
Chan, M. K. M. (2002). Concordancers and concordances:Tools for Chinese language teaching and research. *Journal of the Chinese language Teachers Association, 37*(2), 1-58.
Chandler, B. 1989. *Longman Mini Concordancer*. Harlow: Longman.
Davies, M. (2005). The advantage of using relational databases for large corpora: Speed, advanced queries and unlimited annotation. *International Journal of Corpus Linguistics*, *10*(3), 307-334.
Davies, M. (2009). The 385+ million word corpus of contemporary American English (1990–2008+): Design, architecture and linguistic insights. *International Journal of Corpus Linguistics*, *14*(2), 159-190.
Davies, M. (2010). More than a peephole: Using large and diverse online corpora. *International Journal of Corpus Linguistics*, *15*(3), 412-418.
Hardie, A. (2012). CQPweb—combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics*, *17*(3), 380-409.
Henriksen, B. (2013). Research on L2 learners' collocational competence and development–a progress report. Retrieved from http://www.eurosla.org/monographs/EM02/Henriksen.pdf
Hockey, S. (1988). *Micro-OCP* (*OCP Version 2*). Oxford: Oxford University Press.
Jafarpour, A. A., Hashemian, M., & Alipour, S. (2013). A corpus-based Approach toward teaching collocation of synonyms. *Theory and Practice in Language Studies*, *3*(1), 51.
Kaye, G. (1990). A corpus-builder and real time concordance browser for an IBM PC. In J.Aarts & W. Meijs (Eds.), *Theory and Practice in Corpus Linguistics* (pp.137-162). Amsterdam: Rodopi.
Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of Euralex 2004* (pp.105-116). Bretagne, France: Université de Bretagne-Sud.
Lewis, M. (Ed.). (2000). *Teaching collocations. Further developments in the lexical approach*. Hove, England: Language Teaching Publications.
Li, S. (2016). A corpus-based analysis of collocation errors by American learners of Chinese and its implication for the teaching of vocabulary. *Chinese as a Second Language -The Journal of the Chinese Language Teachers Association-US (CSL)*, *51*(1), 62-78.
Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics*, *24*(2), 223-242.
Palmer, F. R. (1981). *Semantics – A New Outline*. Cambridge: Cambridge University Press.
Rayson, P. (2008). From Key words to Key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519-549.
Reed, A. (1978). *CLOC User Manual*. Birmingham: University of Birmingham.
Scott, M. (1996). *WordSmith Tools.* Oxford: Oxford University Press.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, J. M. (Ed.). (2004). *How to use corpora in language teaching* (Vol. 12). Amsterdam: John Benjamins Publishing.

Souza Hodne, L. C. (2009). *Collocations and teaching: Investigating word combinations in two English textbooks for Norwegian upper secondary school students* (Unpublished master's thesis). University of Bergen, Bergen, Norway.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of language*, *2*(1), 23-55.

Sun, H., & Huang, C. (1998).The distribution characteristics of collocations in texts. *Proceedings of 1998 International conference on Chinese Information Processing*, Beijing, China (pp. 230-236). [孙宏林, 黄昌宁. (1998). 词语搭配在文本中的分布特征. *1998 中文信息处理国际会议论文集(pp. 230-236)*.]

Wu, W. & Li, S. (Eds.). (2009). *Linguistics and Chinese as a Second Language*. Hong Kong University Press, Hong Kong.

Xu, J., & Jia, Y. (2013). The design and development of the R-Gram-based corpus analysis software PowerConc. *Computer-assisted Foreign Language Education*, *1*, 57-62. [许家金, 贾云龙. (2013). 基于 R-gram 的语料库分析软件 PowerConc 的设计与开发. *外语电化教学, 1*, 57-62.]

Xu, J., & Wu, L. (2014). Web-based fourth generation corpus analysis tools and the BFSU CQPweb case. *Computer-assisted Foreign Language Education*, 5, 10-15,56. (许家金, 吴良平. (2014). [基于网络的第四代语料库分析工具 CQPweb 及应用实例. *外语电化教学, 5,* 10-15, 56.]

You, L., & Wang, S. (2005). Rules and distributions of Chinese verb-verb collocations. *Computer Engineering and applications, 23,* 179-181. [由丽萍, 王素格. (2005). 汉语动词——动词搭配规则与分布特征. *计算机工程与应用, 23,* 179-181.]

