Supporting Collocation Learning and Teaching with a Chinese Collocation Profile Database
(建立汉语搭配语料库，促进汉语搭配教学)

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

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

Abstract: Recent studies on collocation have indicated that L2 collocation competence is a crucial factor that distinguishes L2 learners from fluent native speakers. However, mastery of collocation has proved difficult because of the sheer number of collocations in the targeted language. Although a great number of ICT tools have been integrated into language teaching and learning, academically sound and pedagogically enriched computer assisted collocation learning environments are still rare and inadequate in the field of Teaching Chinese as a Foreign Language (TCFL). This paper reports an attempt to construct a Chinese Collocation Profile Database (CCPD) with more targeted selections of collocations, an improved database, and more pedagogically sound online activities in collocation learning. The study first identifies commonly confused collocations for CFL learners from the HSK corpus, and then obtains each collocation's high-frequency collocates from the BCC corpus based upon the strictly applied set of criteria using the corpus analysis toolkit AntConc. Collocation patterns for each collocate are also summarized for learners and teachers. Five types of activities are designed accordingly to train learners on methods for improving upon their collocation competence.

摘要：近年来的搭配研究表明，第二语言学习者的词汇搭配能力是他们有别于流利的母语使用者的关键因素。然而，由于目标语言中搭配的数量非常庞大，第二语言学习者学习并掌握搭配的难度也相当大。尽管目前的语言教学和学习采用了很多的 ICT 工具，但在对外汉语教学领域中，真正以科学的学术研究和合理的教学方法为基础的计算机辅助汉语词汇搭配学习系统还非常少见。本文介绍了构建汉语词汇搭配数据库（CCPD）的理论依据和实现方法。CCPD 对于搭配的选择将更有针对性，所使用的语料库数据也更为准确，针对搭配模式所设计的练习也更符合认知规律。CCPD 的设计的第一步是确定 HSK 动态作文语料库中汉语学习者经常混淆的搭配，然后根据严格设定的统计标准，使用语料库分析工具 AntConc 从北京语言大学 BCC 语料库中提取高频搭配词语。在此基础上笔者再对每个高频搭配词语的搭配模式进行总结和描写。在这些信息的基础上，CCPD 设计了五种类型的练

习，以帮助提高汉语学习者的词汇搭配能力。

Keywords: L2 collocation competence, collocation acquisition, Chinese collocation database, TCFL

关键词：二语搭配能力，搭配习得，汉语词汇搭配数据库，对外汉语教学

1. Introduction

In recent years, collocation research has received wide attention as it relates to the field of second language education. A large number of studies have suggested collocational knowledge is a crucial factor that distinguishes second language learners from fluent native speakers (Palmer, 1933; Hornby, 1974, Hill, 1999, Nation, 2000, Mueller, 2011). Marton (1977) identifies that the incorrect use of collocations constitutes a significant percentage of errors committed by second language learners. Nation (2000) points out that language knowledge is in essence collocational knowledge and the process of learning words requires knowledge of their collocates. Wray (2002) and Nesselhauf (2003) contend that the mastery of collocation is of great importance for second language learners who strive for a high level of competence in a second language, as it helps to enhance both accuracy and fluency.

2. Literature Review

In the field of Teaching Chinese as a Foreign Language (hereafter TCFL), many scholars have identified the acquisition of vocabulary as one of the most challenging aspects for Chinese as a Foreign Language (hereafter CFL) learners (Luo, 1997; Jiang, 1998), and that vocabulary research should play a leading role in the research of TCFL (Zheng, 2005). Research on the learning and teaching of Chinese collocation has received much attention recently, and many studies have been conducted on collocation from multiple perspectives (Fang, 2002; Fu, 2010; Liu, 2010; Zhou, 2007). Taking "Verb + Noun" collocations as an example, Fu (2010) points out that typical patterns of nominal collocates should be identified and taught first to CFL learners, so they can learn more efficiently. Liu (2010) also discusses the need for identifying the patterns of the most frequently used collocations and for creating a corpus-based collocation profile database. Such fundamental work, as she notes, would greatly benefit both teachers and students. With such a set of useful collocations at their disposal, teachers will no longer have to rely on intuition; instead, they will be able to select and create a list of level-appropriate high- frequency collocates for their students and design more effective practice activities. The database will also be able to help draw students' attention to those collocation patterns in context and will help develop their language sensitivity as a result. By studying high-frequency collocates, learners are more likely to generalize to a more abstract pattern and hence increase their ability to produce a more native-like language.

Recognizing the importance of collocation, and the role collocations play in the production of second language, second language teachers have successfully taught collocations by using the corpus linguistics approach. For example, Shin and Nation (2008) argue that determining what to learn and teach remains a challenge as there are so many collocations in English. They believe that identifying the most useful English collocations is an efficient way to improve the language fluency of ESL students. Using the ten million word BNC spoken section as the data source, the study generates a list of the most frequently used collocations in spoken English. As they stated, the list could be useful for teaching elementary speaking courses, and could also serve as the starting point for syllabus design. To address the lack of research in computer assisted collocation acquisition and the limitations of current online collocation learning tools, Wu, Franken, & Witten (2010) proposed a scheme for supporting the learning of English collocation with a digital library. Their design is based on "the psychological conditions that facilitate acquisition: noticing, retrieval and generation" (p. 23). Using a digital library collection prepared by teachers in advance, the design allows learners to search, study, and collect collocations that they have noticed or want to learn through an interface that is specifically tailored for their ease of use. Using this digital library, learners also have the opportunity to expand their knowledge by studying naturally occurring collocations that appear in texts retrieved from corpora such as the British National Corpus or live web data. As the two authors report, the library was tested by student users and the results show that their knowledge of collocations was enriched in a new and engaging way (p. 24). Apart from the aforementioned research and scheme, several useful collocation tools are also available for learners of English, including JustTheWord, COCA, Tango, and the Gutenberg Collocation Tool.

In the field of TCFL, researchers have proposed some computer assisted tools or projects for helping CFL learners expand their collocation knowledge. For example, Chen, Wu, Yang & Pan (2014) report their design of a Chinese collocation retrieval tool that can help CFL learners and teachers search for collocations. The tool is entitled ICE (Intelligent Collocation Engine) and is based on a large part-of-speech-tagged Chinese news corpus. The authors report that the tool was tested by both CFL learners and in-service CFL teachers. The results show that users can successfully find proper Chinese collocates for a given noun, and teachers see this as a useful tool in preparing their teaching materials. Using three corpora including Chinese internet, Lancaster Corpus of Mandarin Chinese, and Corpus of Business Chinese, Sharoff (2006) also created an online automated search engine which can be used to search for Chinese collocations. In addition, Zheng (2005) also proposes to construct a CFL learner-focused lexicon that is both explicit and descriptive in terms of the lexical interdependence. He exemplifies his design principle by providing a compilation sample of the micro structure of the word 解, including its free, fixed and syntactic combination forms with other words. He points out that the future direction for vocabulary instruction in TCFL should be the integration of both a bilingual and learner-focused dictionary as this is an ideal way to provide learners with standardized and prefabricated information (p. 227). He further elaborates that it is imperative for TCFL research to be more dependent on the integration of descriptive linguistics, corpus linguistics and computational linguistics in order to make the discipline a more modernized and scientific one.

3. The significance of the study

Despite the body of research in current literature and the design of computer assisted collocation learning tools in the field of TCFL, there is still much scope for further research and development. Taking the aforementioned factors into account, this paper reports an attempt to construct a Chinese Collocation Profile Database (hereafter CCPD) with more targeted selections of collocations, improved database capabilities, and a more pedagogically sound exploitation of technology in collocation learning. Its design is based on four elements that help to develop the reliable approach to describing Chinese collocation.

Firstly, CCPD is based on the analysis of the learner corpus HSK (*Hanyu Shuiping Kaoshi*, Chinese language proficiency test) Corpus (hereafter HSK Corpus), therefore it targets the challenges and difficulties learners of Chinese encounter in their language production. Nesselhauf (2003) points out that although researchers have made some suggestions on teaching collocations, which of the vast number of collocations in a language should be taught and the manner in which they are to be taught remains unclear (p. 223). He then further suggests that it is crucial to identify the challenges that learners have with collocations. In the field of TCFL, a few scholars also point out the most confusing and frequently misused words should be the core of learner-focused dictionaries, and such a vocabulary list should be based on statistical analysis of inter-language corpus (Zhang, 2008). Although Zhang's (2008) discussion is on the issue of distinguishing synonyms, the principle applies to the creation of a collocation database or dictionary compilation as well. Indeed, the traditional method of compilation of Chinese dictionaries is not sufficient enough to meet the increasing demand of CFL learners' actual learning needs. For example, traditional dictionaries usually provide little explanation about the usage of words, such as context, connotations, pragmatic and register. However, such information is crucial for CFL learners. For this reason, collocations collected for this database are the most frequently misused and confused ones retrieved from HSK Corpus. The detailed procedure of extracting erroneous collocations is discussed in the procedures section.

In addition, the collocation lists and examples that CCPD offers for teachers or learners is based on analyzed and processed data instead of raw data extracted directly from large scale corpora or live web data. While current collocation tools such as the digital library proposed by Wu *et al*. (2010) and the ICE proposed by Chen *et al*. (2014) are useful for both teachers and learners to efficiently search for proper collocates in authentic language data, learners may be overwhelmed by the sheer number of instances of collocation extracted from various genres of texts. As a result, they may get lost in limited and fragmental information by bogging themselves down into unproductive tangential explorations. Since they are the result of automatic extraction, some instances may also not be real examples of collocations and therefore may confuse CFL learners.

In order to tackle this issue, CCPD uses a dataset with various types of possible collocations firstly extracted from the BCC Corpus and then analyzed and selected according to clearly defined criteria. It is hoped that this approach will be more accurate

and efficient in helping students to notice, learn and reproduce collocations in their learning process.

CCPD also seeks to describe the collocation patterns of the most commonly misused collocations for both CFL leaners and teachers. Without careful analysis and systematic descriptions of the frequent collocates, learners who search for collocations using a collocation tool may only be able to notice individual high frequency collocations while failing to see the big picture, which includes the collocation patterns of certain lexical items. These patterns have been argued by some scholars to be more beneficial for learners to learn collocations (Ellis 2013). Lexicological studies in recent years have shown that the importance of the prefabricated lexical chunks and the high-frequency patterns in second language learning process (Sample, 2014). For example, when discussing the theory of priming, Emmott (1997) argues that since speakers or writers often use certain word combinations repeatedly in discourse, listeners or readers may grow to expect the word sequence in text and co-text. He therefore believes that readers may have pre-fabricated patterns in their mind which make the reading comprehension process much faster and more effective. Hoey (2005) further develops the priming theory and claims that "collocation is fundamentally a psychological concept" (Hoey, 2005, p. 7), and each word sequence has a semantic association at a more abstract level. As Hoey explains, in English, the word *hour* is semantically associated with the pattern NUMBER-hour-JOURNEY, which is summarized on the basis of a corpus-based analysis and can be exemplified by the following instances (taken from Hoey, 2005, p. 16):

> *thirty*-hour ride, *half*-hour drive, *four*-hour flight, *two*-hour trip, *three*-hour journey

Hoey (2005) argues that "every word is primed to occur with particular other words" and "every word is primed to occur with particular semantic sets" (p. 13). In other words, each word in language is pre-fabricated in abstract patterns, which are particularly useful for second language learners. According to Hoey (2005), such patterns are like shortcuts which may help speed up the second language learning process. Compared with native speakers, who have been exposed to such patterns in their everyday life, second language speakers have far less opportunities to encounter such high-frequency patterns in their language teaching materials and outside of their classroom. Hence, providing "repeated instances of a word sequence, collocational observations and illustrations" provide a means of shortcutting the language learning process. (p. 185).

An initial analysis of the erroneous collocations in HSK corpus indicates that CFL learners' lack of knowledge about collocation patterns may contribute greatly to error. For example, Li's (2016) study reports that some learners of Chinese have difficulties using 丰富 and 丰盛 with collocate to produce correct collocations. Using BCC Corpus as a native speaker reference corpus, the study obtains the following high-frequency collocates of the two words 丰富 and 丰盛:

Table 1: The collocates and frequencies of 丰富 and 丰盛

Collocates	Frequency	Collocates	frequency
丰盛的晚餐	467	丰富的货物	5137
丰盛的早餐	281	丰富的经验	2792
丰盛的午餐	225	丰富的维生素	1318
丰盛的年夜饭	105	丰富的文化	1002
丰盛的菜肴	87	丰富的资源	993
丰盛的食物	66	丰富的内涵	963

By studying their high-frequency collocates, certain collocation patterns can be summarized. Table 1 shows that (1) what high-frequency nominal items in particular can go with 丰盛 and 丰富 (2) what kind of nouns in general can go with these two words. For example, 丰盛 goes with 晚餐 / 早餐 / 午餐 as high-frequency collocates, but in general, it goes with nouns in the semantic field of MEAL, which would include some low-frequency items such as 酒席、宴席、大餐. So drawing attention to the general pattern is just as important as teaching the individual items. Similarly, 丰富 has its own collocates such as 经验 and 文化，which, by contrast, are abstract nouns. 丰富 also has some collocates such as 货物 and 维生素, which are concrete nouns but without meaning of MEAL. Therefore, it can be said that drawing attention to the general pattern is as important as teaching the individual items.

Lastly, CCPD provides some collocation activities for learners on the basis of sound pedagogical considerations. Lewis (1993) points out that pedagogical chunking should be a frequent classroom activity, and authentic language materials should be provided for learners to experience, analyze, generalize, and experiment with lexical chunks. Studies on collocation have also suggested that three aspects should be taken into consideration when teaching collocation: awareness raising, retrieval, and production (Nation, 2000). He further suggests that learners should first be made aware of the most frequent and immediately useful collocates. Using authentic data extracted from the BCC corpus, CCPD aims to provide a variety of activities that help learners of Chinese notice, remember, and use collocations appropriated in various contexts. Detailed activity demos are illustrated in the CCPD design section.

Thus, by using analyzed controllable data and targeting at the challenges CFL learners face in their collocation learning process, CCPD aims to develop a more systematic approach to providing CFL learners with a more accurate, efficient and engaging way to learn collocations. The tool also emphasizes the importance of offering a variety of pedagogically tuned learning activities, which have proven effective in helping learners notice collocations, use them in their language production, as well as internalize them and add them into their lexical reservoir.

4. The proposed CCPD

4.1 The Definition of Collocation in this study

In the current literature, researchers have defined collocations from various perspectives, but in general, these definitions can be classified in two groups: phraseological view or statistical view based on frequency of co-occurrence. For example, Firth (1957) defines collocation as statements of the habitual or customary places of certain words. Halliday and Hasan (1976) regard collocation as linear co-occurrence of certain lexical items that have significant proximity in terms of syntagmatic relation. Adopting a corpus linguistic approach, Sinclair (1991) defines collocation as the occurrence of two or more lexical items within a short space of each other in a text. In this study, we adopt a definition similar to that of Benson, Benson, & Ilson (1986) and that of Wei (2001). In their definition, Benson, Benson, and Ilson define collocation as "certain words combined with other words or grammatical constructions" (p. ix). They further define such constructions as recurrent, semi-fixed combinations and classify them as grammatical and lexical collocations. Wei (2001) also divided such combinations into three categories: free combination, restricted combination (collocations), and fixed combination (idioms). This study focuses on lexical collocations. Table 2 shows some examples of the erroneous collocations in the list.

This study first identifies high frequency combinations in the BCC Corpus through statistical procedure, and then includes and excludes specific word sequences according to an analysis of the word combinations identified. Detailed procedure is reported in the procedures section.

Table 2: Examples of commonly confused and misused collocations in HSK Corpus

Misused collocations	Correct collocations suggested by HSK Corpus
*危害人权	侵犯人权
*丰富的晚饭	丰盛的晚饭
*社会里	社会上
*调试心情	调节心情
*在他的眼光里	在他的眼里
*健康地生长	健康地成长
*提高生产	提高产量
*用力学习	努力学习

4.2 The two corpora used in this study

The study uses two corpora - the HSK Corpus and the BCC Corpus. The HSK Corpus is developed and maintained by the Research Center for Studies of Chinese as a Second Language at Beijing Language and Culture University. The HSK Corpus collects about 11,600 essays (approximately 4.3 million Chinese characters) written by learners of Chinese for the HSK test. It not only contains important text attributes such as nationality, gender, and age of CFL learners who took the test, but also provides a wealth of inter-language information such as the statistical data of characters, vocabulary, sentences, and discourse contained in compositions. Various types of learner errors are also tagged

in this corpus, which makes it possible for researchers to extract them by using its online search engine at http://202.112.195.192:8060. The following examples explain how the coding system works for tagging incorrect use of vocabulary in the HSK Corpus:

(1) 在路上闻不到烟的味，看不到烟蒂{CC 烟根}。

(2) 吃过丰盛{CC 丰富}的晚饭后，我们一家人就坐在饭厅里长谈到深夜。

In sentences (1) and (2), CC stands for any erroneous words, wrong choice of words, self-made words, and erroneous collocations. The word in Sentence (1) 烟根 is a self-made word, the correct word should be 烟蒂, while in Sentence (2) the word 丰富 is a wrong collocate of 晚饭, the correct collocate should be 丰盛。Therefore the wrong words are put in the curly bracket after the symbol CC, and the suggested correct words are placed before the first curly bracket.

The study also uses the BCC Corpus as a L1 reference corpus. It was developed by the Institute of Big Data and Education Technology of Beijing Language and Culture University. The corpus has a 15 billion character collection of samples of present-day written language from a wide range of sources such as microblogging, science and technology, literature, and the press. Similar to the HSK Corpus, the BCC Corpus also has an online search engine interface at http://202.112.195.249/bcc/. The online concordancer also offers a statistical function, which not only allows any user to search collocations using some formulaic expressions, but also provides statistical data of frequencies of different collocates.

4.3 The procedures

The procedure of the establishment of CCPD can be illustrated as in Figure 1 (c.f. next page).

Since CCPD aims to provide collocation patterns of the most commonly misused collocations, the first step is to extract all the erroneous production of collocations from the HSK Corpus using its online concordance tool. By searching CC as a string type, 49178 tokens of CC were returned. The next step is to manually exclude instances that are not erroneous collocations. For example, the word 烟根 in Sentence (1) is a self-made word, not a wrong use of collocation. Therefore, it is removed from the list. This process proves to be time-consuming due to the sheer number of instances extracted from the HSK Corpus. The end result of this process is to produce a list of the most commonly confused and misused collocations (hereafter CMC) in the HSK corpus, which is then categorized into the various types of collocations as shown in Table 3:

Table 3: Collocations types in HSK Corpus

Collocation types	Examples
Verb + Noun	侵犯人权
Adjective + Noun	丰盛的晚餐
Number ＋ Measure word＋Noun	一只蝴蝶

Noun ＋Location Noun	社会上
Preposition＋Noun＋Location Noun	在他的眼里
Adjective + Noun	健康地成长
Gerund + Noun	创业精神
Noun + Noun	工薪阶层
Noun + Adjective	竞争激烈

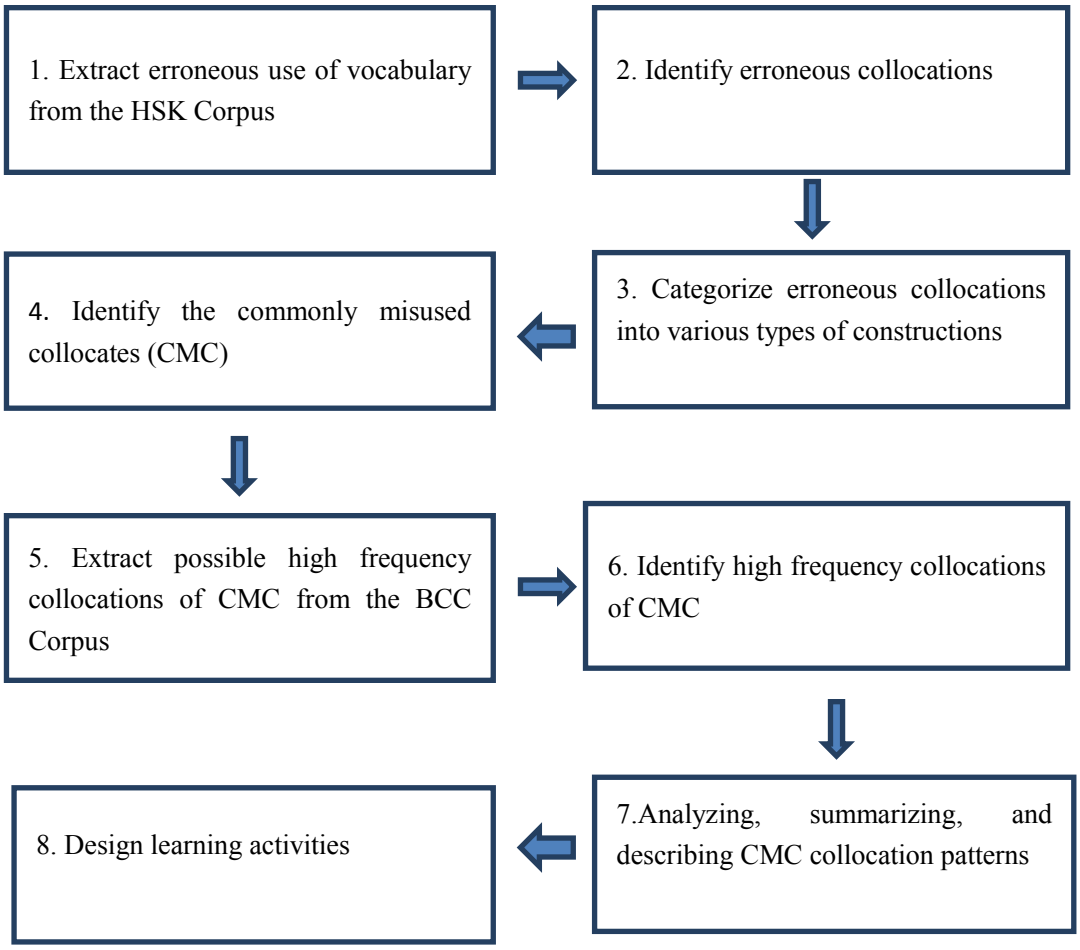

Figure 1: The procedure of establishment of CCPD

Once the CMC list and the categorization of CMC are completed, the next step is to extract possible high frequency collocations of CMC from the BCC corpus. Because collocations are domain and/or genre specific (Hoey, 2005, pp. 9), this study only uses one of BCC's sub-corpora - literature sub-corpus - as its source data to ensure that the text properties are relatively homogeneous. Future development of CCPD may include more types of text such as newspapers and magazines, scientific and technical text, etc.

Extracting instances of possible collocations of a CMC is relatively straightforward by using the online search tool offered by BCC. For example, the authors typed 丰盛的 N (N stands for noun) into the search box and chose literature sub-corpus.

2213 instances were returned and downloaded as a text file, which were then further processed for the identification of high frequency collocates using AntConc (Windows 3.4.4). The process of the text file includes two steps: firstly, the <u> and </u> marks surrounding the word"丰盛的 N (such as 晚餐/大餐)" are removed. The Chinese words in the file are then parsed using ICTCLAS (Institute of Computing Technology, Chinese lexical Analysis System), and then the parsed text is saved as a .txt format with the encoding code set as UTF-8.

The next step is to use AntConc to perform a collocation analysis. Drawing on previous studies on the identification of Chinese collocations using a statistical approach (Bai, 2004; Sun, 1998; You, 2005, and Wang, 2006), this study sets the span of words as (-3, +4) for analyzing verbs, (-2,+1) for nouns, and (-1,+2) for adjectives, and uses the three most commonly used statistical measures - frequency, MI value, and T-score - to determine whether two words co-occur by chance or if they are lexically primed to collocate with each other. This study adopts Wang's (2006) approach, setting the two measures as MI ≥3, T≥2.33, if the MI value and T score of a certain co-occurrence meet the condition, then it can be potentially regarded as a typical and most commonly used collocation. Below is a screen capture of the analysis result of the collocates of 丰盛：

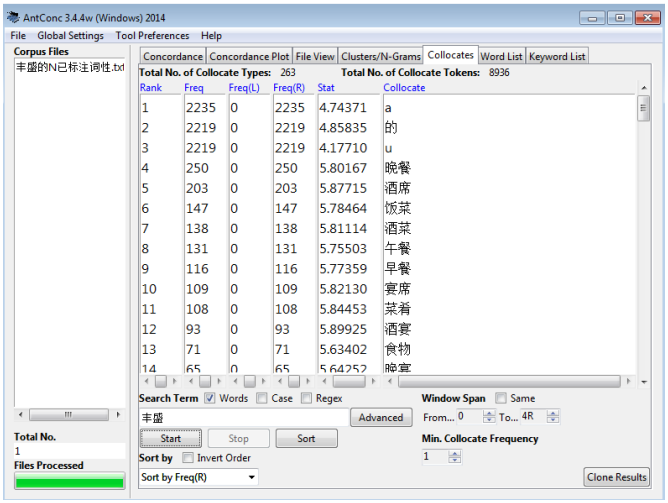

Figure 2: The screenshot of the analysis result of the collocates of 丰盛

The analysis results is then exported and input into an excel sheet, Table 4 shows the list of collocates that can enter into the collocational construction 丰盛+的+N.

Table 4: Results of AntConc collocation analysis (T≥2.33)

Frequency	T-Score	Collocates	Frequency	T-Score	Collocates
257	15.75165	晚餐	21	4.49844	礼物
204	14.04104	酒席	21	4.49143	美食
147	11.90441	饭菜	20	4.38233	宴会
139	11.58136	酒菜	18	4.17449	午宴
134	11.36629	午餐	17	4.04129	食品

118	10.66758	早餐	13	3.54763	酒食
114	10.49653	宴席	13	3.53426	早饭
110	10.30888	菜肴	12	3.40382	家宴
93	9.48207	酒宴	10	3.11148	嫁妆
72	8.33193	晚宴	10	3.1064	年夜饭
71	8.25647	食物	10	3.01496	地方
56	7.35666	筵席	7	2.60325	美餐
49	6.87149	大餐	7	2.60325	早点
35	5.79932	晚饭	7	2.59111	犒劳
33	5.64948	酒筵	7	2.49396	美味
33	5.62152	佳肴	6	2.40358	祭品
26	4.95095	招待	6	2.39702	礼品
26	4.93205	款待	6	2.39047	酒饭
22	4.59452	午饭	6	2.39047	宴请

Table 5: Results of AntConc collocation analysis (M≥3)

Frequency	MI	Collocate	Frequency	MI	Collocate	Frequency	MI	Collocate
252	5.813	晚餐	18	4.402	款待	4	4.500	饮食
204	5.884	酒席	18	5.960	午宴	4	4.960	草地
147	5.784	饭菜	17	5.655	食品	4	5.638	膳食
138	5.811	酒菜	13	5.960	酒食	4	5.152	羊肉宴
132	5.766	午餐	13	5.660	早饭	4	5.375	海鲜
117	5.785	早餐	11	5.719	家宴	4	5.960	年饭
109	5.857	菜肴	10	5.822	年夜饭	4	5.152	伙食
109	5.821	宴席	10	5.960	嫁妆	4	5.375	人生
93	5.899	酒宴	10	4.423	地方	4	5.960	中餐
71	5.634	食物	7	5.960	美餐	3	5.544	自助餐
69	5.728	晚宴	7	5.960	早点	3	3.223	生活
55	5.858	筵席	6	5.375	酒饭	3	5.544	爱情
48	5.737	大餐	6	5.737	祭品	3	4.375	烤肉
35	5.663	晚饭	6	5.544	礼品	3	4.737	成果
33	5.916	酒筵	5	5.960	饭食	3	5.960	夜餐
33	5.544	佳肴	5	5.696	酒肴	3	5.544	夜宵
22	5.612	午饭	5	3.638	美味	3	5.960	供养
21	5.651	美食	5	5.960	筵宴			
21	5.767	礼物	5	5.474	猎物			
20	4.727	招待	5	5.960	果实			
19	5.564	宴会	5	5.960	席面			

The list of collocates are then further checked to achieve accuracy. For example, in both Table 4 and Table 5, the word 地方 is listed，although its T and MI value are

high,丰盛的地方 doesn't seem to make much sense. By checking its context, the authors found that 地方 is actually used in a structure like 物资最丰盛的地方/水草较丰盛的地方. Here, 地方 is not modified by 丰盛，but by 物资最丰盛/水草较丰盛 instead. 地方 should therefore be removed from the collocate list.

Based on this result, the collocation pattern of 丰盛+的+N is then summarized. This is done by analysing the semantic features of all collocates in Table 4 and 5 and identifying the similarities between them. In Table 4 we can see that there are 33 nouns that are in the semantic field of [MEAL], which represents 89% of the total number of nouns. Similarly, in Table 5, 87% (50 out of 57) of the nouns have the semantic property of [MEAL].

The last step of the procedure is to design activities for learners of Chinese. The goal of these activities is to help learners gain a degree of familiarity with the collocations presented in CCPD, and consequently develop an awareness and sensitivity to them. Through these activities it is hoped that they can eventually acquire the collocational competence and add these important collocations into their reservoir. The following sections describe the proposed learning activities in detail and the pedagogical considerations behind them.

4.4 Collocation activities

Once high-frequency collocates of CMC are identified and CMC's collocation patterns are summarized, the next step is to design learning activities. Before going into detail about the activities, however, one important pedagogical principle which needs to be noted here is the concept of *noticing*. In the current literature on collocation learning, many studies have pointed out that noticing is crucial for L2 learners to learn collocations (Lewis, 2000; Nation, 2000). By noticing, learners are made aware of the existence of collocations and therefore pay close attention to them as part of the language rather than as part of a message. Lewis (2000) also emphasizes that if learners can focus explicitly on some aspect of linguistic form of the input (for example collocation), it helps to accelerate their acquisition process. (p. 160). By noticing high-frequency collocations, learners are encouraged to think beyond the words, and realize that collocations are pre-constructed chunks that are at the disposal of native speakers during communication. This is a key to achieving collocation competence (the awareness of collocations and the capacity to use them) that differentiates these learners from native speakers. Although noticing is of great importance, collocations are often ignored by L2 learners, as Lewis (2000) warned, "don't assume students are noticing collocations and recording for themselves, they won't unless you train them to" (p. 163). The learning activities proposed here, therefore, mainly focus on raising CFL learners' awareness of high-frequency collocations. Take the word 树立 as an example. The CCPD provides a collocation pattern for learners (see Figure 3). The pattern page lists the high-frequency collocates, and also describes the main semantic feature of these collocates, namely that they are all in the semantic field of [+ABSTRACT]. The CCPD also provides some sentence examples that these collocates are used in (see Figure 4). These examples are

also purposely used here to raise their awareness that collocations (such as 树立…观念，树立…意识，树立…榜样) are similar to pre-constructed chunks, and that often there are only certain nouns that can enter into such pre-construction. The sentences presented in CCPD are all taken from the BCC corpus, and minor changes are made to better suit CFL learners. On the basis of these patterns and examples, various types of activities are then designed to train learners to internalize and absorb such collocational constructions into their mental lexicon by noticing and practicing.

Figure 3: The collocation pattern of 树立

Figure 4: Sentence examples of high-frequency collocates of 树立

Drawing on insights from these studies and the collocation digital library designed by Wu *et al*. (2010), the CCPD includes five types of exercise: gap filling, matching, multiple choice, error correction, and re-writing.

(1) Gap filling activity

This type of exercise is in essence another form of presentation for the collocation patterns of certain CMC. Learners are given some words from a high-frequency list of a certain CMC and are asked to choose the right collocates to complete each collocation. Figure 6 demonstrates one such activity, which focuses on finding the right noun for the verb 树立. The purpose of this exercise is to reinforce the collocational construction learners have just noticed on the collocation pattern and examples pages.

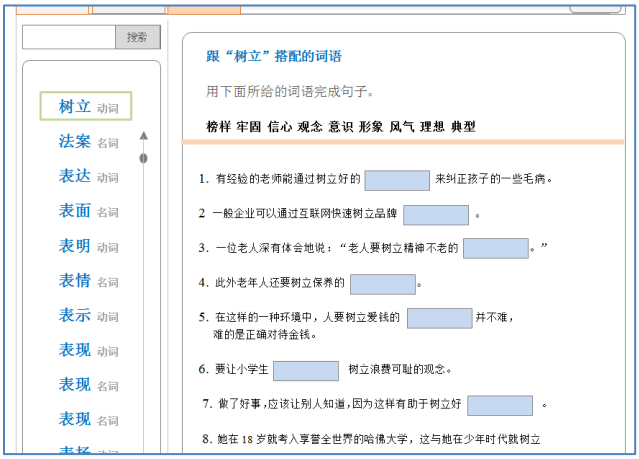

Figure 5: Gap fill exercise example of 树立

(2) Matching

Matching is to some extent similar to gap filling activity, however, CCPD uses this activity more to help learners differentiate groups of words that share similar meanings or commonly confused words that have different collocates. For example, 选拔 and 选择 are two similar commonly confused words (Liu, 2010). Using the aforementioned procedure of finding high-frequency collocates, the authors identified the high frequency collocates of the two words from the BCC literature sub-corpus, Table 6 lists some of these (the list is not complete due to the space limitation).

Table 6: Collocates of 选择 and 选拔

（选择+）N			（选拔+） N		
职业	目标	时间	人才	官员	官吏
对象	道路	方案	干部	贤能	贤才
机会	地方	人才	人员	将领	士兵
方向	方法	专业	接班人	英才	高手

It can be seen that apart from the word 人才, 选择 and 选拔 actually do not share their collocates. It is therefore useful for learners to notice their high-frequency collocates respectively by drawing attention to this difference.

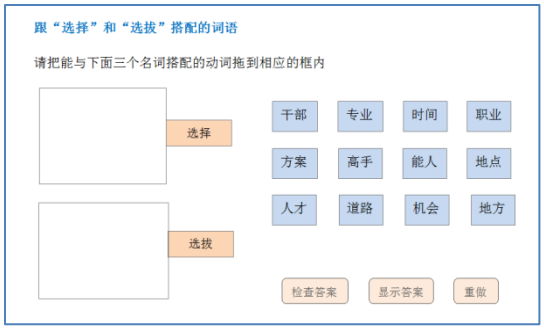

Figure 6: Matching exercise example of 选择 and 选拔

(3) Multiple choice

The multiple choice activity is very similar to the gap filling exercise, but offers more limited choices for learners. This exercise serves well when attempting to differentiate collocates of synonyms. For example, Li (2016) points out that 丰富 often collocates with words such as 经验, 知识, 信息, which are all in the semantic filed of [+ABSTRACT], and words such as 资源,维生素,食物, which are in [+CONCRETE], while 丰盛 often collocates with words such as 晚餐,早餐,午餐,年夜饭,饭菜,食物, which are all in the semantic filed of [+CONCRETE,+MEAL]. The following example shows that multiple choices can be used to focus learners' attention explicitly on such differences.

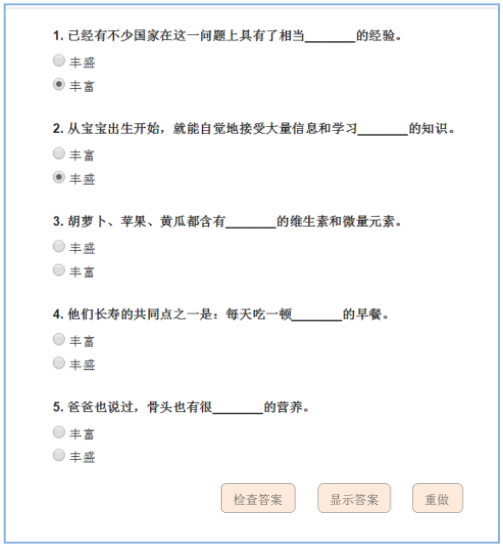

Figure 7: Multiple choices example

(4) Error Correction

This activity is designed to check if learners have grasped the collocational constructions after they study the collocation patterns of certain CMC and their actual use in the sentence examples provided in CCPD. The erroneous sentences are all taken from

the HSK corpus and some minor changes are made to suite L2 learners. Compared with the above three activities, correcting errors is relatively more difficult, as learners have to first read and identify which group of words are collocations and then decide which of them are wrong. By providing learners with such erroneous collocations, they hope to consolidate their internalization of collocational constructions and hence help them achieve a greater acquisition of collocations.

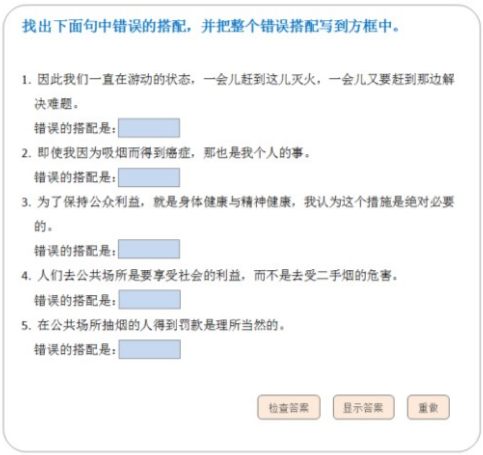

Figure 8: Error correction example

(5) Rewriting

Unlike the above four activities, which focus learners more on acquiring receptive knowledge, rewriting activity targets more at learners' productive knowledge. As Wu *et al*. (2010) discuss, repetition and use are two effective ways to help learners remember a collocation (p. 7). When learners use a collocation to produce a sentence or construct a conversation, they firstly have to retrieve it from their lexical repertoire, and then have to decide if the collocation is semantically and pragmatically appropriate for that circumstance. This process is very productive as it not only consolidates the receptive knowledge (what they noticed about the collocation) but also encourages them to productively control the collocation in a specific context.

Figure 9 (c.f. next page) shows an example of such activity. It firstly presents a paragraph of text with a few high-frequency collocations. In this example three collocations can be identified: 得出… 结论, 提高…水平, 承担…成本. Depending on learners' language proficiency level, the collocates 结论, 水平, and 成本 can either be underlined or not underlined. Once these collocates are identified, learners are required to use these collocations to rewrite the whole paragraph. For this exercise, the following short paragraph can be provided as a reference answer:

> 最近的一项研究得出这样一个结论，适度的低出生率可以提高一个国家里人们的生活水平，因为出生的人口少了，每个家庭所承担的养孩子的成本就降低了。

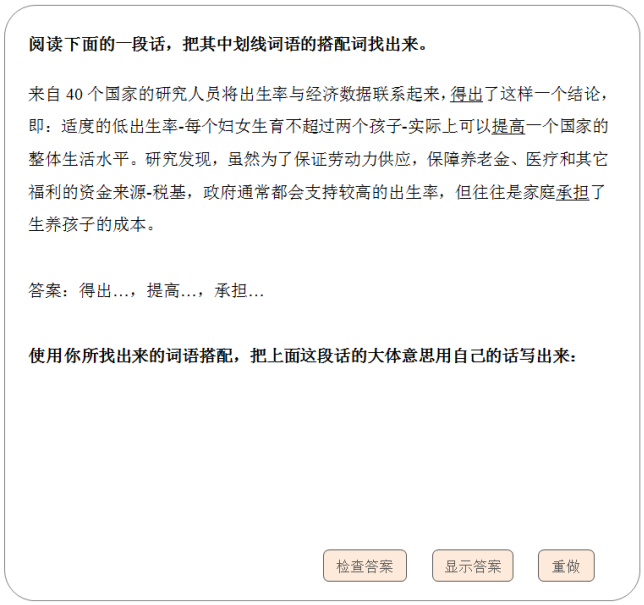

Figure 9: example of rewriting activity

Learners can also be required to underline the collocations used in their writing again, which further enhances the awareness of these pre-constructed chunks.

5. Limitations and Future development

One limitation that CCPD has is that it only investigates the collocations on the basis of the literature sub-corpus of BCC. Therefore, the high-frequency collocates listed in CCPD only present part of the big picture. Future development can include more sub-corpora of BCC. High-frequency collocates of certain CMC in different text genres can also be compared and analyzed, and the result may be helpful for teachers and learners to raise the awareness of genre-specific collocations.

Another limitation may lie in the lack of design present in the automatic activity and answer generation in what Wu *et al*. (2010) created. In their proposal, an exercise design interface is provided for teachers and learners to create exercises at different levels of linguistic difficulty, as well as to make the activities more collaborative and competitive. Since the goal of establishing CCPD is to form a foundational database of Chinese collocations, offering more in-depth knowledge such as collocation patterns and a high-frequency collocate list of each collocate, the manual processing and analysis involved in this project are essentially necessary and time consuming. It is currently not viable to design an exercise design interface similar to that in Wu *et al*. (2010). Such a design could be beneficial for both teachers and learners if it can be accomplished in the future development of CCPD.

6. Conclusion

This study presents a design of a Chinese Collocation Profile database for supporting CFL learners' development of collocation competence. Drawing on current studies on L2 collocation acquisition, the study first identifies commonly confused collocations for CFL learners in their writing for HSK. Once the list has been generated, the high-frequency collocates of each collocation component from the list are then identified using the corpus analysis toolkit AntConc and based on carefully applied criteria. The high-frequency collocates provide a solid foundation for the authors to carry out semantic analysis of the collocation patterns for each collocate. Such patterns are crucial to enable CFL learners to firstly learn a collocation pattern by studying high-frequency items, and then generalize to a more abstract pattern. To draw learners' attention to these high-frequency collocates and their patterns, CCPD provides five types of activities to help learners notice and practice collocations in context, with an aim to raise awareness regarding collocations as pre-constructed lexical chunks in language, and to increase their familiarity with the linguistic features of collocations. Among these types of exercises, gap filling, matching, multiple choice, and error correction are rather receptive, while rewriting activity is more productive and engaging.

The description of the CCPD in this study is only a theoretical attempt. More practical issues are to be taken into consideration when constructing the actual online learning environment. A demo website has been built to demonstrate the ideas discussed in this study. The URL for the demo site is http://ccpd.space/.

References

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.]

Benson, M., Benson, E., & Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam/Philadelphia: John Benjamins.

Chen, H. H. J., Wu, J. C., Yang, C. T. Y., & Pan, I. (2014). Developing and evaluating a Chinese collocation retrieval tool for CFL students and teachers. *Computer Assisted Language learning*, *29*(1), 21-39.

Ellis, N. (2013). Construction grammar and second language acquisition. In T. Hoffmann & G. Trousdale (eds.), *The Oxford Handbook of Construction Grammar*, 365-378. Oxford: Oxford University Press.

Emmott. C. (1997). *Narrative Comprehension: A discourse Perspective*. Oxford: Clarendon Press.

Fang, Y. (2002). On Lexical Collocation and the Teaching of Chinese as a Foreign Language. *Journal of Lianyungang Vocational and Technical College, 15*(3), 58-61. [方艳. (2002). 论词语搭配与对外汉语教学, *连云港职业技术学院学报*, *15*(3), 58-61.]

Fu, N. (2010). A multi-perspective analysis of the semantic features of nouns in verb-noun collocations. *Journal of Ninxia University (Humanities and Social Sciences*

edition), *32*(6), 57-59. [付娜. (2010). 动名词搭配中名词语特点的多层面分析. *宁夏大学学报（人文社会科学版）*, *32*(6), 57-59.]

Firth, J. R. (1957). Modes of meaning. In J. R. Firth (Ed), *Papers in linguistics* 1934-1951 (pp.190-215). Oxford: Oxford University Press.

Halliday, M. & Hassan, R. (1976). *Cohesion in English*. London: Longman.

Hill, J. (1999). Collocational Competence. *English Teaching Professional*, 11, 3-6.

Hoey, M. (2005). *Lexical Priming: A new theory of words and language.* London: Routledge.

Hornby, A. S. (1974). *Oxford Advanced Learners' Dictionary*. Oxford: Oxford University Press.

Jiang, X. (1998). The Study on vocabulary acquisition and its implications to the teaching. *Language Teaching and Research,3*, 65-73. [江新. (1998). 词汇习得研究及其在教学上的意义. *语言教学与研究*, 3, 65-73.]

Lewis, M. (1993). *The lexical approach*. Language Teaching Publication, England.

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 Collocational Errors by American Learners of Chinese and its Implication for the Teaching of Vocabulary. *Journal of The Chinese Language Teachers Association* , *51*(1), pp. 51-69.

Liu, F. (2010). A corpus-based study on lexical collocation and the teaching of vocabulary in TCFL. *Modern Chinese-Language Study*, 6, 115-117. [刘凤芹. (2010). 基于语料库的词语搭配研究与对外汉语词汇教学. *现代语文：下旬. 语言研究*, 6, 115-117.]

Luo, Q. (1997). Psychological characteristics of English learners' advanced vocabulary learning process and corresponding teaching strategies, *the proceedings of the 5th International conference on Chinese language pedagogy,* Beijing: Peking University Press. [罗青松. (1997). 英语国家学生高级汉语词汇学习过程的心理特征与教学策略. *第五届国际汉语教学讨论会论文选.* 北京：北京大学出版社.]

Marton, W. (1977). Foreign vocabulary learning as problem no.1 of language teaching at the advanced level. *Interlanguage studies Bulletin*, *2*(1),33-57.

Mueller, C. M. (2011). English Learners' Knowledge of Prepositions: Collocational Knowledge or Knowledge Based on Meaning?. *System: An International Journal of Educational Technology and Applied Linguistics*, *39*(4), 480-490.

Nation, P. (2000). *Learning vocabulary in another language*. Cambridge University Press.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. Applied linguistics, *24*(2), 223-242.

Pamler, H. E. (1933).*Second Interim Report in English Collocations*. Tokyo: Kaitakusha.

Sample, M. G. (2014). An Overview of The Lexical Approach And Its Implementation At A Public Elementary School In South Korea. *Journal of International Education Research (JIER)*, *10*(4), 271-278.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. *WaCky*, 63-98.

Shin, D., & Nation, P. (2008). Beyond single words: the most frequent collocations in

spoken English. *ELT Journal*, *62*(4), 339-348.

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

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*.]

Wang, L. (2006). Quantitative approach to collocation analysis. *Journal of Shanghai Normal University (Philosophy and Social Sciences edition*), 6, 117-122. [汪腊萍. (2006). 词项搭配的定量分析方法. *上海师范大学学报（哲学社会科学版）*, 6, 117-122.]

Wary, A. (2002). *Formulaic Language and the Lexicon*. New York: Oxford University Press.

Wei, N. (2001). *The Definition of Lexical Collocation and Its Research Framework*. Shanghai: Shanghai Jiaotong University Press. [卫乃兴. (2001). 词语搭配的界定与研究体系. 上海: 上海交通大学出版社.]

Wu, S., Franken, M., & Witten, I. H. (2010). Supporting collocation learning with a digital library. *Computer assisted language learning*, *23*(1), 87-110.

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.]

Zhang, B. (2008). On confusable words in Chinese interlanguage and its research methods, *Language Teaching and Research,* 6, 37-45. [张博. (2008). 第二语言学习者汉语中介语易混淆词及其研究方法. *语言教学与研究*, 6, 37-45.]

Zheng, D. (2005). On the construction of CFL learner-oriented contrastive lexicon. *Proceedings of the 8th International conference on Chinese language pedagogy.* [郑定欧. (2005). 谈对外汉语学习型对比词库的构建. *第八届国际汉语讨论会论文选*. 北京: 世界汉语教学学会.]

Zhou, X. (2007). *The Study on lexical collocation and TCFL* (Unpublished doctoral dissertation). Shanghai International Studies University, Shanghai, China. [周新玲. (2007). 词语搭配研究与对外汉语教学(博士论文). 上海外国语大学, 上海, 中国.]

