A Corpus-Based Study on High Frequency Verb Collocations in the Case of "HAVE"

Xiujuan Zhou
Qingdao Harbor Vocational and Technical College, Qingdao, Shandong, China
Email: zhou848122@163.com

[Abstract] On the basis of a corpus-driven approach, this research investigates high-frequency verb collocations in the case of *have* by Chinese non-English major learners. Results show that despite the most frequent use of the verb *have*, the learners make use of relatively low collocation types. The learners tend to simply overuse the words related to the topic or given by the writing directions. It is also found that in the use of *have* collocation, the learners are inclined to be affected by mother tongue interference and overgeneralization.

[Keywords] high-frequency verbs, collocation, HAVE, corpus

Introduction

Collocations have received increasing attention in language in recent years. They are important and play an essential role in language learning. Corpus-based studies both at home and abroad have offered further insights into collocations. For example, it is found that verb+ noun collocations are frequent and among the most difficult for the learner (Howarth, 1996). In China, statistics based on the Chinese Learner English Corpus (CLEC) have shown that verb+ noun collocation makes up the greatest number of all the types of collocational errors (Gui & Yang, 2003; Wang, 2005).

Among countless verbs, high-frequency verbs are particularly worthy of mention. High-frequency verbs are verbs that are frequently used and, consequently, turn up early in frequency lists. As revealed by many researches, high-frequency verbs are not only frequent and important, but also tend to be problematic for foreign language learners (Altenberg & Granger, 2001; Hasselgren, 1994). The present study intends to give some insight into the use of high-frequency verb collocations by Chinese English learners. More precisely, the study concentrates on one major representative of this group of verbs, the verb *have.*

The analysis is based on a learner corpus we constructed of one million words. Through detailed analysis, we first aim to get a general view of the features of *have*+ noun collocation. Then the second aim is to identify the typical errors and non-errors in *have*+ noun collocation and to trace some underlying factors related to collocational misuses. Finally, the pedagogical implications are to be discussed.

The study is designed to address the following questions: 1) What is the pattern of *have*+ noun collocation? 2) What are the characteristics of the learners' use of *have*+ noun collocation? 3) What factors result in the *have*+ noun collocation errors?

Review of the Literature

Definitions of Collocations

The term "collocation" is used widely different and often rather vague senses in linguistics and language teaching. Firth first brought the idea of collocation into prominence. According to Firth (1957), "you shall know a word by the company it keeps" (p.12). His definition of collocation is: "collocations are actual words in habitual company" (p. 99). After Firth, there have been two main views on the notion of "colloccations." One view is that collocation is the co-occurrence of words at a certain distance; from which the name "statistically oriented approach" (Herbst, 1996, p. 380) or the "frequency-based approach" is derived (Nesselhauf, 2004). Sinclair is a typical representative of the frequency-based approach. He defines collocation as "the occurrence of two or more words within a short space of each other in a text" (1991, p. 170).

In the other view, collocation is considered as a type of word combination, most commonly as one that is fixed to some degree, but not completely. This view has been called the “significance oriented approach” (Herbst, 1996, p. 380) or the “phraseological approach” (Nesselhauf, 2004). Cowie, a main representative of this view, considers collocations as “a composite unit which permits the substitutability of items for at least one of its constituent elements” (1981, p. 224) with the other elements being constant.

It is considered that there are two basic kinds of collocations: grammatical collocations and lexical collocations. Benson et al.’s (1997) classification is that the former consists of a dominant word (noun, adjective/participle, verb) and a preposition or a grammatical construction, and the latter refers to co-occurrence of two lexical elements. They distinguished lexical collocations into six types: “verb+ noun,” "adjective+ noun,” “noun+ verb,” “noun+ noun,” “adverb+ adjective,” and “adverb+ verb.” In the present study, what is discussed is the verb+ noun collocation. In addition, collocation in the study is established on the basis of statistics by use of corpus evidence, as was done by Sinclair.

Learner Corpora and the Analysis of Learner Language

Learner corpora can be defined as “systematic computerized collections of texts produced by language learners” (Nesselhauf, 2005, p. 40). It has some crucial advantages over other kinds of data in the field of second language acquisition (SLA). Since its existence, it has greatly facilitated the analysis of learner language and language teaching. Two of the well-known learn corpora abroad are the Longman Corpus of Learners’ English (LCLE), and the International Corpus of Learner English (ICLE). The famous learner corpora in China are the Chinese Learner English Corpus (CLEC), Spoken and Written English Corpus of Chinese Learners (SWECCL), to name but a few.

Error analysis (EA) was a traditional way to analyze learners’ errors. With the advent of learner corpora, a new approach to the analysis of learner errors won its popularity. That is computer-aided error analysis (CEA). Equipped with the methods and tools of corpus linguistics, CEA is more powerful and is favored over EA. It is the research methodology adopted in this study.

Based on CEA, there have been a number of studies on high-frequency verb collocations in learner language. Chi *et al.* (1994) used the Hong Kong University of Science and Technology (HKUST) to analyze the combinations with the verbs *have, make, take, do* and *get.* They classified learners’ errors into two types: confusion of the five words with each other and confusions of the five verbs with other verbs. Altenberg and Granger (2001) examined English as a foreign language (EFL) learner use of the verb *make*. Results show that EFL learners, even at an advanced proficiency level, have great difficulty with a high-frequency verb *make*. (p. 173). Deng’s (2004) study focuses on the collocation patterns of six delexical verbs used by Chinese EFL learners at two different proficiency levels. The results indicate that the Chinese learners show a strong tendency of overuse of these verbs.

To sum up, previous research findings suggest that learners of English, including Chinese EFL learners, have collocational problems with high-frequency verbs.

Data and Methodology

The Learner Corpus Used

The learner corpus consists of 633 compositions written by non-English major students in Dalian Maritime University. When the writings were collected, the students were in their second year of college study and would take the CET-Band 4 exam soon in May 2006. The compositions were collected as assignment writings in class. Students were required to finish writing of given topics within class as a writing exercise. However, they didn’t know that their compositions would be used for research, so the validity and reliability of the corpus was promised.

All the essays collected are narrative and argumentative, amounting to 103,027 words. The topics are 1) “Cheating in College,” 2) “Water Shortage,” 3) “Computer and Human Brain,” 4) “Traffic Jam,” 5) “Going to College is Expensive,” 6) “Travel,” 7) “Money is the Root of All Evils,” and 8) “Violence on TV.”

CLEC is a widely recognized corpus which is claimed to represent overall proficiency levels of Chinese EFL learners. In CLEC, ST3 stands for tertiary non-English majors Band 4 proficiency level. Here in our own compiled corpus (for convenience, we name it DLMU), materials were collected just before CET-Band 4, so it is comparable to ST3 in CLEC. To see whether data in the corpus we constructed is valid, a comparison between DLMU and CLEC is given in Table 1.

Table 1. Descriptive Data of the Corpora

Corpora	Type	Token	Number of texts	Average length
DLMU	6050	103,027	633	163
ST3(CLEC)	6658	216615	1317	164

The table shows that although the tokens and number of texts in ST3 are two times more than DLMU, they are much the same in terms of types and average text length. Thus, our corpus DLMU is appropriate to represent the characteristics of Chinese EFL learners. One thing worth mentioning is that compared with CLEC, the size of the corpus in our study may be small. However, something can still be revealed, and this is, in Granger's (1998) view, "especially true for learner language, which is an extremely heterogeneous variety of English" (p.146).

Data Retrieving and Analysis

To obtain the data necessary for the present study, a number of FoxPro programs were applied. The following is to tell the selection of data for analyzing:

1) By running a WORDLIST.prg FoxPro program (see Appendix 1), we get the basic data of the corpus: types, tokens, and most important of all, word frequencies. The purpose of this step is to choose the most frequently used verbs for investigation.
2) Identify every instance of the chosen verbs respectively from the corpus. To obtain the concordance lines, a concordance FoxPro program package KEYWORD5.exe (the core program see Appendix 2) was used in this step. After running the program, the item being studied (keyword or node) is put at the center of each line and all the co-occurrence word types are in a span of 4 words. The search of KWIC (Key Word in Context) is the basis of the methodology for the present study.
3) Count the noun collocates of each chosen verb manually and compute the co-occurrence frequency of each word type before they are saved in the file DATA.txt.
4) Test the extracted verb+ noun collocability. Three measures will be adopted in the study: frequency, likelihood ratios, and Z-score.

To count the frequency of joint occurrence of the verb and each noun collocate is the simplest way to pick out the typical collocations that our learners have used. However, this method alone may present a biased measure of word associations. Likelihood ratios, as the name suggests, tell us how much more likely one collocation is than the other. Collocation is about degrees of likelihood. Likelihood shows words that are statistically weighted preferences. Here in the present study, it is calculated with the aid of the third FoxPro program LIKELIHOOD.prg (see Appendix 3) to assess significance of collocates. The likelihood ration is calculated with the following formula:

The likelihood ratio = log(P)×C12+log(1-P)×(C1-C12)+log(P)×(C2-C12)+log(1-P)×((N-C1)-(C2-C12))-log(P1)×C12-log(1-P1)×(C1-C12)-log(P2)×(C2-C12)-log(1-P2) ×((N-C1)-(C2-C12))

P: probability (P = C2 / N, P1 = C12 / C1, P2 = C2 - C12 / N-C1)

C: number of occurrences (C1 for the keyword, C2 for the collocate, C12 for the co-occurrence of the words)

Z-score is another statistic that we used to observe the collocability between lexical items in order to ensure the reliability of the results. Normally, a z-score of 2 or higher can be considered to be significant.

Based on the data obtained in LIKELIHOOD.prg, the fourth FoxPro program ZVALUE.prg (see Appendix 4) was run to calculate the Z-score. The formula for calculating Z-score is as follows:

$P = C2/N \quad E = P \times M \quad M = (2S+1) \times C1 \quad SD = \sqrt{p(1-p)M}$

$Z = (Cj - E) / SD$

N: Size of corpus C1: frequency of key word M: the mini-text S: span of context
C2: frequency of collocate in corpus Cj: frequency of C2 in M
E: expected number of collocates in M

Verbs to be Investigated

It is assumed that common verbs, i.e. frequently used ones, are more worth studying than uncommon ones. Thus, in order to find which the often used verbs are in the DLMU learner corpus, a frequency list needs to be referred to. According to the first step, the word frequencies of DLMU are calculated, and the top eight most frequently used verbs are listed in Table 2.

Table 2. Top 8 Verbs in the Learner Corpus

Rank	Verb	Frequency	Rank	Verb	Frequency
1	have	1026	5	think	378
2	do	541	6	use	348
3	make	441	7	get	315
4	go	395	8	take	313

The eight verbs in the table are among the 15 most frequent verbs in any corpus-based list of high-frequency verbs (Altenberg & Granger, 2001:173). The study chooses the verb *have* for investigation, not only because it tops the list of DLMU corpus, but also because it is a "lexical teddy bear" (Hasselgren, 1994) often used by Chinese EFL learners. As far as the verb *have* is concerned, we should take into consideration that *have* can serve as both main verb and auxiliary verb. As the current study only focuses on its collocational use, occurrences of the verb with auxiliary uses are beyond the scope of our investigation. In order to be accurate, instances like the following examples are excluded manually:

[204] that the traffic jams HAVE been a big problem
[208] In recent years there HAS been a marked increase
[299] how many of them HAVE considered the expense of
[387] the TV program producers HAVE found the key to

Results and Discussion

After the last step of data collection, the typical collocations of the chosen verb *have* were obtained. Some of the results are presented in Table 3, and more results are displayed in Appendix 5 for reference. In the table, all collocates in Column W2 are ranked in the decreasing order of the likelihood ratios, and at the same time the equivalent Z-score is also provided.

Table 3. Top 19 Collocates of Have in Decreasing Order of Likelihood

Rank	W1	W2	C12	C1	C2	Likelihood	Z-score
1	have	money	94	581	1201	340.64	4.26
2	have	Side	20	581	32	165.60	14.42
3	have	advantage	13	581	32	91.90	8.93
4	have	Time	24	581	353	77.74	1.44
5	have	choice	11	581	40	67.39	6.30
6	have	ability	8	581	16	60.87	7.98
7	have	Idea	10	581	47	55.49	4.93
8	have	access	5	581	6	46.43	8.51
9	have	opinion	10	581	112	37.47	1.81
10	have	emotion	5	581	11	36.73	5.94
11	have	effect	6	581	24	35.41	4.33
12	have	experience	7	581	43	34.77	3.26
13	have	chance	6	581	33	31.21	3.34
14	have	friend	7	581	63	29.25	2.13
15	have	disadvantage	5	581	26	26.60	3.20
16	have	ambition	3	581	4	26.60	6.21
17	have	confidence	3	581	4	26.60	6.21
18	have	affect	5	581	35	23.45	2.42
19	have	dream	5	581	36	23.16	2.35

As can be seen from the above table, collocate *money* occurs strikingly frequently in learners' writings: 94 times. Its likelihood ratio is on the top of the rank list. It should be noted that the word *money* is rather theme-related. Its concept is closely related with one of the writing topics, "Going to College is Expensive," and the word itself is directly in the topic "Money is the Root of all Evils." For the excessive use of *money*, a related process is probably involved in students' writing, such as, "title recycling" in Nesselhauf's (2005) terminology. Title recycling refers to "the use of words or expressions that occur in the title or topic given by the teacher" (p. 148). Here in our study, it is the re-use of the single word *money* in the title. It can be seen that learners often rely on words they have just encountered in the title or topic in their essays. Learners may assume that by re-using words in the title or topic, their writings could be theme-related.

The collocate *side* co-occurs 20 times with *have* in the corpus. Its likelihood value is ranked second, but its Z-score is the largest. That is to say, the collocability of *have* and *side* used by learners is quite strong. To go a bit deeper into the high, frequent use of the collocation with *side*, we examine all the specific instances which learners used. The following concordances show how students used the *have+ side* collocation pair.

[456] convenient however, every coin HAS its each two sides.

[469] shadow of money. Everything HAS its two side, people

[470] But as every coin HAS its two sides, money

[471] all evils. However everything HAS its two sides. If

[472] old saying, every coin HAS its two sides. Maybe

[473] bad master. Each coin HAS its two sides. Money

[658] we can’t let coin HAS only one side. Every

[974] As the coin HAS two effect sides, it

[977] it However, every win HAS two sides .No one

[978] and the evil. Money HAS two sides .one side

[979] Looking at it .It HAS two sides and each

[980] is ordinarily. Every coins HAS two sides, cheating is

[981] But a win always HAS two sides, the development

[982] are . But everything HAS two sides, there are

[983] before. But every coin HAS two sides. As you

[984] carried out. One coin HAS two sides. I think

[985] their howes. Every coin HAS two sides. Make good

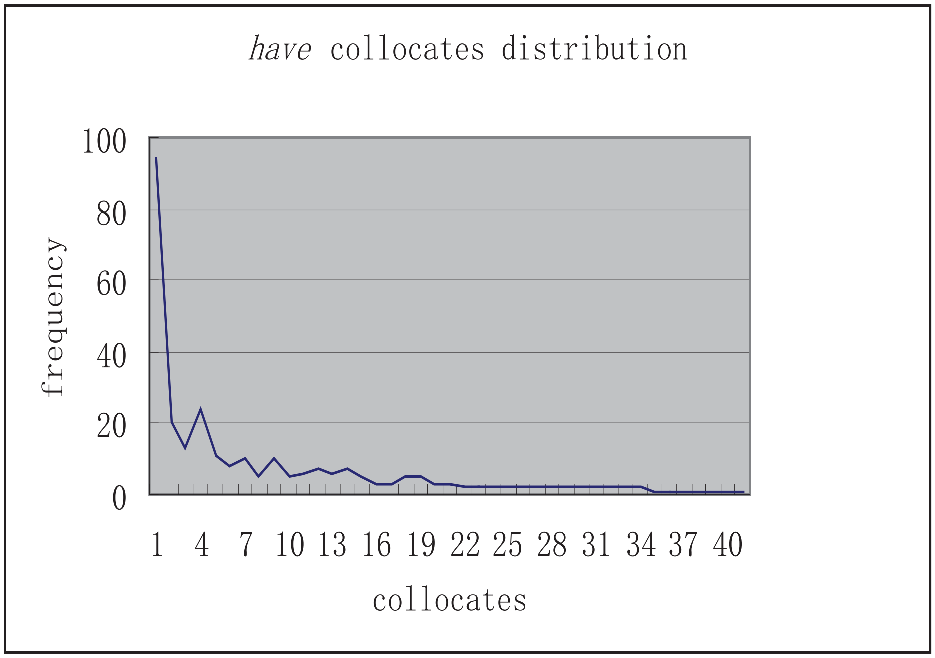
[986] computer. However, every coin HAS two sides. The computer

[987] others things. The coin HAS two sides. So we

[988] student, farmer, government. coin HAS two sides. Some others

What is striking is that in these instances *have* and *side* occur in almost identical stretches of text. Learners all employed the phrase "two sides of the same coin," and, in most cases, they referred to money. This shows learners’ strong tendency of repetition in their writing. When the same or similar concept is to be expressed several times, learners simply repeat an expression.

Possible reasons for this phenomenon are that learner is convinced that the collocation is appropriate for the concept in question and repeats it in order not to take any risks or that the expression is familiar to learners. In writings, they tend to rely on what is learned early or familiar. However, from a general point of view, frequencies of the noun collocates of *have* in DLMU corpus are sparsely and unevenly distributed. It can be clearly seen from the table listed in Appendix 5. *Have* collocates with the word *money* most frequently, co-occurring as many as nearly 100 times. Noun collocates *time* and *side* are ranked second and third, with 24 and 20 times individually. There is a sharp and obvious decline from collocate *money* to *time* and *side*. What’s more, in general, the co-occurring frequency of noun collocates is low. Verb h*ave* collocating with most of the nouns less than 10 times. The following graph illustrates the distribution of the collocates clearly.



Among 148 noun collocates of the verb *have*, there are only 5 items with more than 10 co-occurrences, and 3 items occurred 10 times. The number of co-occurrences with less than 10 times reaches 143. The reason might partly be due to the small size of the corpus. However, on the other hand, it may indicate that learners only use very few *have*+ noun collocations.

The verb *have* is particularly productive in combining with a following noun phrase to form a large number of different collocations. A careful examination on the noun collocates suggests that more than half of the nouns used by the learners have a direct translation equivalent in Chinese. For example, in Table 3, *have money, have time, have advantage, have disadvantage, have ability, have idea, have experience, have chance*, etc. are all thought to be equivalent to their Chinese counterparts in meaning. As is known, L1 plays an important part in learners' collocation production. It is assumed that if the concept the learners had in mind can be expressed in L2, learners are likely to produce large numbers of collocation and use it as "lexical teddy bear." However, L1 influence or transfer can be both positive and negative, while positive transfer can result in an acceptable expression, such as the examples listed above. Negative transfer would yield deviant collocations. The following examples illustrate some typical errors committed in the learners' writings:

[403] the college students' brains HAVE good development and they

[827] of human's activities. Human HAVE thousands of years' development,

[432] like a crime which HAS high risk but high

(correct form: run/take risk)

[517] bad manner. And it HAS many harms. For one

(correct form: cause/do harm)

[727] people. Conversely, the travel HAVE some harm, as well,

(correct form: cause/do harm)

[824] quiet big shame to HAVE this deed as a

(correct form: do/perform)

[958] car rate. The roads HAVE too much burden without

(correct form: bear/carry)

Collocations of *have* in the above examples all have corresponding expressions in Chinese. However, they are deviant collocations. Examples [403] and [827] are the combination of *have* and *development.* Looking up the *Oxford Collocations Dictionary* (2003), we can find that the noun *development* is typically followed by the verb *occur* or the phrase *take place*, rather than collocating with *have*. It's clear that students made Chinglish mistakes in these two examples. Errors in the rest of the examples are direct misuse of verb in the collocation: *Have* should be replaced by another verb. The correct form is given below each individual example in the above. In these collocations, L1 exerts a strong negative influence. The errors result from students' literal translation from their mother tongue to the target language, i.e. English. Apparently, *Have* is being overused by students. It can be seen as an effect of overgeneralization of the target pattern.

Conclusion

To conclude, there are two major findings concerning the use of *have* in our study. First, despite the overall low co-occurrence with *have*, there are some noun collocates occurring with high frequency. *Money* and *side* are two fairly extreme instances, which can be explained as a result of repetition and title recycling. Second, as to learners' collocational performance, *have+* noun collocations were often produced correctly but were also found to be deviant particularly often. The most important factor for its collocation difficulty is that learners employed a literal translation strategy and borrowed L1 equivalence of a collocation in the production of L2. In other words, learners tend to look for the corresponding L1 expression to express what they wish to express in the L2. This can sometimes lead to deviation because of negative transfer. Deviations of *have+* noun collocation are largely an effect of overgeneralization. Learners cling to *have* like a "lexical teddy bear," for it is learned early, and, in their minds, the verb is "widely usable and above all safe" (Hasselgren, 1994, p. 250).

These results have several practical pedagogical implications. First, like Altenberg and Granger's (2001) study of *make*, although *have* is a high-frequency verb, learners still have difficulty in producing appropriate collocations. To improve learners' use of high-frequency verbs, an increase of exposure to their typical collocations is needed. Teachers should raise learners' awareness of the complexity of high-frequency verbs. For example, given that L1 interference is apparent in *have+* noun collocations, teachers could point out the collocational divergences between the two languages. Second, to avoid repetition and title recycling in learners' writings, teachers could teach a number of collocations that occur frequently in connection with this topic before the essay is written.

Based on our own constructed learner corpus, the present study mainly investigates the collocation use of the high-frequency verb *have*. Because the corpus size is comparatively small, learners' use of collocations needs to be analyzed on the basis of greater amounts of data. More research is also necessary for other high-frequency verbs for a comprehensive understanding. This paper is only a tentative study. It is hoped, however, that it can throw some light on the use of high-frequency verbs by Chinese EFL learners.

References

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22,* 173-194.

Bahns, J. (1993). Lexical collocations: A contrastive view. *ELT Journal*, *47,* 56-63.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL learners collocation? *System*, *21*, 101-114.

Benson, M., Benson, E., & Ilson, R. (1997). *The BBI dictionary of English word combinations.* Amsterdam: John Benjamins.

Chi, M. A., Wong, P.K., & Wong, C. M. (1994). Collocaional problems amongst ESL learners: A corpus-based study. In L. Flowerdew & A. Tong (Eds.), *Entering text* (pp. 157-165). Hong Kong: University of Science and Technology.

Corder, S. P. (1967). The significance of learners' errors. *International Journal of Applied Linguistics*, *5,* 161-170.

Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics, 2,* 223-235.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *System*, *26,* 163-174.

Deng, Y.C. (2004). Collocation patterns of delexical verbs in Chinese EFL learners' writing. (Master thesis, Dalian Maritime University, 2004).

Dulay, H., Burt, M., & Krashen, S. (1982). *Language two*. Oxford: Oxford University Press.

Ellis, R. (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Firth, J. (1957). *Papers in linguistics.* London: Oxford University Press.

Gass, S., & Selinker, L. (Eds.). (1983). *Language transfer in language learning.* Rawley: Newbury House.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Ed.). *Phraseology: Theory, analysis, and applications.* (pp. 145-160). Oxford: Oxford University Press.

Gui, S.C. & Yang, H.Z. (2003). *Chinese learner English corpus*. Shanghai: Shanghai Foreign Language Education Press.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4,* 237-260.

Herbst, T. (1996). What are collocations: Sandy beaches or false teeth? *English studies, 77* (4), 379-393.

Howarth, P. (1996). Phraseology in English academic writing. *Some implications for language learning and dictionary making.* Tübingen: Niemeyer.

Kennedy, G. (1998). *An introduction to corpus linguistics.* Harlow: Longman.

Nesselhauf, N. (2004). What are collocations? In D. Allerton, N. Nesselhauf, & P. Skandera (Eds.). Phraseological units: Basic concepts and their application (pp. 1-21). Basel: Schwabe.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Odlin, T. (1989). *Language transfer*. Cambridge: Cambridge University Press.

Oxford Collocations Dictionary for Students of English (2003). Oxford: OUP.

Sinclair, J. (1991). Corpus, concordance and collocation. Oxford: Oxford University Press.

Wang, H.H. (2005). Investigating the collocational behavior of Chinese EFL learners: A corpus-based approach. (Doctoral dissertation, Shanghai Jiao Tong University, 2005).