CSCanada
Higher Education of Social Science
Vol. 7, No. 1, 2014, pp. 87-91
DOI: 10.3968/5182

ISSN 1927-0232 [Print]
ISSN 1927-0240 [Online]
www.cscanada.net
www.cscanada.org

Analyzing Collocation Errors in EFL Chinese Learners' Writings Based on Corpus

HUO Yanjuan[a],*

[a]School of Foreign Languages of China West Normal University, Nanchong, China.
*Corresponding author.

Received 12 March 2014; accepted 2 June 2014
Published online 28 July 2014

Abstract
English writing, a creative construction process and a crucial way of language output, have been recognized as an indispensable part in EFL (English as a Foreign Language) learning for Chinese students. Based on CLEC (Chinese Learner English Corpus), the present paper conducts a study on collocation errors in the compositions of Chinese students. Although the emphasis will be on the description and analysis of collocation errors, attention will also be given to pedagogical implication by means of studying learners' language with the help of corpora and concordance program. This paper relies on corpus which can provide large and systematic authentic language collection, and tries to investigate the characteristics of Chinese learners' real EFL output.

Key words: collocation; CLEC; BNC; EFL; CET

Huo, Y. J. (2014). Analyzing Collocation Errors in EFL Chinese Learners' Writings Based on Corpus. *Higher Education of Social Science*, *7*(1), 87-91. Available from: URL: http://www.cscanada.net/index.php/hess/article/view/5182
DOI: http://dx.doi.org/10.3968/5182

INTRODUCTION
To Chinese learners, writing is a hard nut to crack. Collocation errors accounting for a lion's share of errors in Chinese EFL learner's written productions have attracted the attention of researchers and teachers. In the past years, we have witnessed a growing awareness of the importance and challenging for students to produce native like collocation. In order to improve the accuracy and fluency of target language output, teachers stress the significant status of collocation in writings and require students to draw wide attention to the native speakers' preference for a certain sequence of words. However, the result is far from satisfactory. It is noticeable that although students are quite familiar with the meaning of each word, the errors of collocation are still everywhere in the students' written form. The present research is based on the data retrieved from Chinese Learner English Corpus (CLEC), which enjoys the high reputation in China by providing reliable and huge number of written samples collected from almost all levels of EFL learners in China. The author intends to provide pedagogical implication in pushing forward the current English teaching by utilizing the advanced corpus and its searching program.

1. THE BASIC THEORIES
1.1 About Collocation
As this paper mainly does some research on the errors of collocation in EFL, it is necessary to make clear what the collocation refers to. Firth, who was regarded as the father of collocation, put forward the conception of collocation. He told us that collocations of a given word are statement of the habitual or customary places of that word in collocation order but not in any other contextual order and emphatically not in any grammatical order（Firth, 1957a). Halliday claimed that lexis seems to require the recognition merely of linear co-occurrence together with some measure of significant proximity, either a scale or at least a cut-off point. It is this syntagmatic relation which is referred to as collocation（Halliday, 1975). From the above saying by scholars, collocation is described as a kind of company relationship, the habitual co-occurrence of individual lexical items. Therefore, we can know a word by the company it keeps, and this co-occurrence of two or more lexical items is mutual expectancy. Lexical items of collocation should not be simply stringed together according to user's free

will but are concerned with grammatical restrictions as well as lexical determination, which mean that collocation partners should have mutual prediction and restriction.

1.2 General Introduction to CLEC

CLEC is an EFL learner corpus project leaded by the prestigious professor Gui Shichun of Guangdong University of Foreign Studies and principal scholar Yang Huizhong from Shanghai Jiaotong University. The corpus is a collection of English compositions of five different groups of Chinese learners ranging from senior middle schools to colleges. There are two kinds of writing in CLEC which are named as free writing labeled as ST2, ST5, ST6 and examination writing conducted in CET4 and CET6 examinations which are labeled as ST3, ST4 respectively. The detailed information is listed in the following （Gui, Yang, & Yang, 2005):

1) ST2 represents the group of students in senior middle school;
2) ST3 represents the group of freshmen and sophomores, most of whom will take CET4;
3) ST4 represents the students of juniors and seniors, most of whom will take CET6;
4) ST5 represents the students of freshmen and sophomores majoring in English;
5) ST6 represents the group coming from juniors and seniors majoring in English;

Moreover, the topics of writing cover a wide range including campus, friends, knowledge, life, English, students' view on society, some phenomena in the world etc. The sampling involves the category or genre of the writing which include descriptive, expository and argumentative essays collected from almost all levels of English learners in China. Thus, the random selection from CLEC can represent the status quo of Chinese students' interlanguage. The data of Chinese EFL learner's writings are all from the search result of CLEC（Gui & Yang, 2003) .

1.3 BNC

BNC (The British National Corpus), containing 100 million wordswords, is regarded as a huge corpus of modern English with a wide range of genres retrieved from spoken and written English. As the best known national corpus, BNC is managed by an industrial/ academic consortium led by Oxford University Press. The BNC comprises the number of w-units (POS-tagged items) as high as 98,363,783 although its orthographic words are just under 100 million. It is very convenient for users to visit http: / /corpus. Byu.edu / bnc /. The BNC website allows users to freely and easily search the 100 million words of the contemporary English (Lou, 2001). About 90% words in BNC are extracted from written texts which has a variety of sources such as regional and national newspapers, specialist periodicals and journals etc.. The left 10 percent — about 10 million in total — are spoken word collection extracted from transcribed speech, recorded in both formal and informal contexts.

2. COLLOCATION ERRORS BASED ON CLEC AND BNC CORPUS

2.1 The Searching Tool

The present study uses AntConc which was released by Laurence Anthony on March, 2006, to retrieve data from CLEC. AntConc is a concordance package comprising three main tools — Wordlist, Concord and File view. According to Laurence Anthony, a concordance program can find and display a huge number of examples in varied contexts and situations quickly and efficiently (Laurence, 2005). AntConc can be free downloaded from the Laurence Anthony Laboratory website. AntConc enables the researcher to get the target words, collocations, sentences immediately through searching CLEC corpus which can be sorted and displayed in various forms.

2.2 General Views of Collocation Errors in CLEC

All the compositions in CLEC are manually error-tagged into 61 types of errors identified in this corpus, running across sentence, phrase and word level. The tagged errors are of significance to this study within which [cc] refers to collocation errors. There are six subtypes of [cc] errors (named cc1, cc2, cc3, cc4, cc5, cc6) as explained in below:

(a) cc1 means improper noun-noun collocation, for example:

> surprise and disease life [cc1,3-]. All this will lead them to a longer life yeaomy in developing countries, the state of health [cc1,-1]gains in those countries has been eople like to do the same job in their work life [cc1,1-]. Because they like their job and good " That often happen [vp3,2-] to everyone student [cc1, 1-]. Haste makes waste. So we don't

(b) cc2 means improper noun-verb collocation, for example:

> young man cry [vp6,6-0] : "my pocket didn't see [cc2,-]", The conductor asked: "Did you the day except the people whose leg can't walk [cc2,3-0]. I didn't buy some [pr6,-] new o go [wd5,3-2] to travel, but the weather rains [cc2,3-0]. I only stay at home. Pases [fm1,-] uld have a test as soon as the holidays went out. [cc2,1-1]. [sn8,s]That meant I would not be

(c) cc3 means improper verb-noun collocation, for example:

> he asked for going hometown [cc3,1-]. The agreed that [sn8,2-]. From then on, no And when you read the truth [cc3,2-], you will wrooy [fm1,-] about the hero and heprecious than money. Let us catch time [cc3,1-] together to study! October 20th Saturday afternoon. Hold this good choice [cc3,3-]. and [fm3,-s] we'll send the five final

(d) cc4 means improper adjective-noun collocation, for example:

> the weather is very hot and you will wear cool [cc4,-1] clothes. The leaves of the trees grow It was a nervous and senseless day

[cc4,4-]. It was completely [wd2,2-2] because of the eit for his office. The emperor was very pleasant [cc4,4-]. Then he gave them a lot of gold, silk awkward. So he thought [vp1,1-3] a rotten idea [cc4,2-]. He give [vp6,d] a [np7,-1] order to

(e) cc5 means improper verb-adverb collocation error, for example:

We jumped, ran, smiled aloud [cc5,2-] in the street. I seemed as if we had not see I read the texts louder [cc5,5-4] as nobody [wd4, 1-2] near me. I was carefu He believed he cheats at all [cc5,5-0]. He "put" on the few clothes and walked down They played in my house a few time [cc5,-]. I though [fm1,-]we were truthful [wd3,-1]frie

(f) cc6 means improper adverb-adjective collocation, for example:

how to kick the shuttlecock [cc6,-1]nearly. But when I studied in the senior schoo it looks like some crystals. That is beautiful [cc6,-2]very much. In the wind, leaves shake two new teachers are so young, and they are nice [cc6,-2]very much. Though they stayed you will see all the thing [np6,2-0] have great [cc6,-1] changed. New shoots of tree comes

With error tags in CLEC, the collocation errors can be identified easily by observing inserted errors labels and square brackets. For instance, in the sentence *Let us catch time [cc3,1-] together to study! October 20th, [cc3,1-]*stands for collocation error between verb and noun, that is to say "catch time" is wrongly collocated. The number stands for how many words should be taken into consideration. *[cc3,1-]* means the verb and noun collocate of the node catch is one word to the left, i.e. time.

The distribution of six types of collocation error in CLEC can be displayed in the following Table.

Table 1
Distribution of Six Types of Collocation Error in CLEC

Error type	cc1	cc2	cc3	cc4	cc5	cc6	Sum
ST2	95	44	213	84	31	22	489
ST3	76	207	598	114	47	14	1056
ST4	95	60	505	163	36	8	867
ST5	28	8	90	50	6	3	185
ST6	69	26	160	61	8	2	326
Sum	363	345	1566	472	128	49	2923

In order to show the distribution of the six types of collocation errors in the five sub-corpus (ST2-ST6) vividly and clearly, the author transforms the Table into the following.

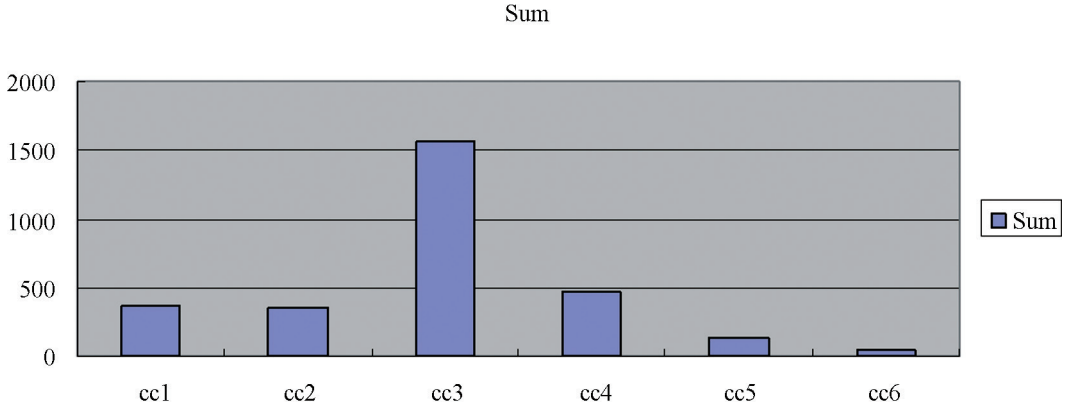

Figure 1
Comparing the Distribution of Collocation Errors

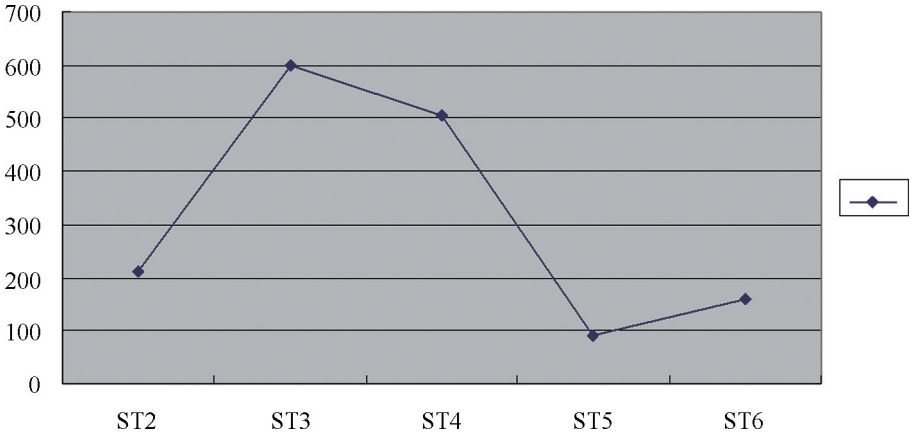

Figure 2
The Distribution of cc3 in CLEC

Table 1 and Figure 1 show that in terms of the total number of collocation errors made by the five groups of students, verb-noun collocation is the most challenging part for students with wrong cases as high as 1,566 while cc6(adverb-adjective collocation) seems the easiest part in the acquisition process. In terms of the student group, ST3 makes the largest number of collocation errors as high as 1,056 cases followed by ST4 (867 cases) and ST5 makes

the fewest errors in collocation (185 cases). In one word, verb-noun collocation (cc3) is the significant error made by students which will be investigated in detail.

2.3 Detailed Discussion

The powerful concordance software Antconc (3.2.1) is used for extracting cc3 errors from the CLEC corpus and the function of Concordance and File View is the main tool to use. Concordance is used to find out all the inappropriate verb-noun collocations (cc3) in the corpora, while File View is exploited to see the whole context in which cc3 occurs.

It can be seen from Figure 2 that the inappropriate choice of the number of cc3 occurring most frequently in ST3 so it deserves our attention to give a careful observation in cc3 by ST3. In China, CET examinations are large-scale standardized tests and enjoy high status in society which score is believed objective and convincing by students and teachers. As introduced in the above part, ST3 stands for writing samples written by Non-English majors in CTE4 (College English Test Band 4 Examination). In order to observe the highest frequency cc3 in ST3, Antconc (3.2.1) is used for extracting examples from CLEC, the following list showing part of the concordance lines with cc3 as the node:

make any social service [cc3,3-] for people, by thus [wd3,1-] , we can get kno the campus, we can learn knowledge [cc3,1-] from books. Although [wd3,-s] , in the societ When we learned to drive [cc3,-1] bikes, it was possible [wd2,1-1] difficult fo For instance, learn [vp4,-s] to play computer [cc3,1-] is also important. We read book only not] it is a [np7,-1] error. Play [vp4,-s] computer [cc3,1-] is necessary for us to practice it [pr1,s-] .people know how to live and take activities [cc3,1-] , how to prolong their lives by [wd3,-2] scieoutside the campus, First [fm3,-] , we can learn [cc3,-2] more trends of economic growth in When we play sports [cc3,1-] , for example, play basket-ball. [sn2,s] That don't be afraid of failure, just make practice [cc3,1-] . And one day you'll find that you are very gour country's economic trends, and I will learn [cc3,-3] the different cultures in the most of people You should take more practice [cc3,2-] so that you can master the important part of ieve "practice makes perfect," you should pay [cc3,-2] more efforts to win it [pr1,s-] . [sn1,s-the world. In [pp2,-1] campus, we can only learn [cc3,-2] the knowledge and theory. But being a...

From the observation of the above examples, we can easily find out that students are used to seek help from verbs which are equivalent to the Chinese verb having similar meaning. For example, they use a lot of the English word "learn" to express the equal Chinese verb "xue xi" because in Chinese, "xue xi" can collocate with a lot of nouns. It is convenient and easy for students to give a literal translation whenever students feel it is beyond their ability to offer a proper expression. Moreover, neglecting the exceptions and restrictions in target language, Chinese non-English major students often generalize the rules and create some deviant collocations on the basis of correct collocations they have learned before. Since the ultimate goal of Chinese students learning English is to achieve native-like English proficiency, it is advisable for teachers to employ the native speaker's corpus BNC serving as standard examples to follow. Taking verb-noun collocation of "play" and "learn" as the verb for example, the nouns which co-occur usually frequently with the word "learn" and "play" within a certain span (usually from 0L to 4R words in span) will be generated. The following Table 2 shows the first 10 nouns collocate with "learn" in BNC corpus.

Table 2
The First 10 Nouns Collocate With "Learn" in BNC

Rank	Nouns	Frequency
1	thing	31
2	English	22
3	language	20
4	skills	18
5	French	10
6	techniques	9
7	German	7
8	lessons	7
9	languages	6
10	music	6

If we want to know more about how to use the phrases, we can click the words in noun column such as thing, or English, language, etc. The following figure is KWIC (keyword in context display).

CLICK FOR MORE CONTEXT [?]

1	KRR	S_interview	A	B	C	we are also very, very efficient. We can ready quickly, we can **learn things** quickly, we can cope, we can act on the whole rather better
2	J9N	S_meeting	A	B	C	No I don't think either erm (pause) The whole emphasis of Homeworks is to **learn things** there which we can then bolt on to the main chain, I'll
3	J9N	S_meeting	A	B	C	no that is not the intention. I do believe, however, we'll **learn things** from it that we can quickly adapt and put into the hundred and seventy
4	FMH	S_classroom	A	B	C	got to learn half of them. Don't want to-- I want you to **learn things** but there's no point in learning everything twice is there or learning more
5	JP8	S_classroom	A	B	C	of your own bedroom, plus (unclear) is a very useful time to try and **learn things**. If you turn over. (SP:JP8PSUNK) (unclear) it was heart searching (unclear) (SP:PS4HA)
6	JT4	S_classroom	A	B	C	, how do we know **things**. Or more specifically, start considering how you **learn things**. (pause) Now you should all have a spare scrap of paper that's
7	K6J	S_classroom	A	B	C	the same as three take way seven. (SP:K6JPS000) Mhm. (SP:PS5LU) You sort of **learn things** like that. Times, that's okay, three times four, the
8	JYN	S_lect_soc_science	A	B	C	help because it, you know, it certainly makes it a lot easier to **learn things**, take information in, revise **things** erm (pause) whatever. (SP:PS6NP) I do
9	JK1	S_tutorial	A	B	C	it's more easily able to erm, observe **things**, take in **things**, **learn things**, and has a lot of advantages over us older, us adults.
10	FXT	S_brdcast_discussn	A	B	C	pedagogical skill. There are so many **things** involved. Obviously people appreciate that children **learn things** and they can help them, but the question is, Why? How
11	KRH	S_brdcast_discussn	A	B	C	which it's operating, and also so that sometimes they may be able to **learn things**, and pick up useful tips and hints about the way to handle a
12	FUE	S_speech_unscripted	A	B	C	at the beginning to get you talking about it. Because only then can you **learn things**. I often say jokingly that (unclear) you make a comment to somebody and
13	AD1	W_fict_prose	A	B	C	' I like to learn,' he said now,' you can always **learn things**.' Mrs Shottery's husband had collected everything, and china figures in
14	AMB	W_fict_prose	A	B	C	you **things** you didn't want to know.' Why do I have to **learn things** I don't want to know?' asked Endill. His mother laughed
15	AMB	W_fict_prose	A	B	C	the other boys left him alone. They thought him odd because he wanted to **learn things**. They liked him even less when they discovered he hated playing rugby,
16	AMB	W_fict_prose	A	B	C	away isn't such a good idea. It would be good to talk and **learn things** from him." Nonsense!' cried the Headmaster.' There

Figure 3
Context of "Learn Things" (cited from BNC website)

BNC offers a great number of sentences which are the context of the verb-noun collocations equipped with detailed source information including date, title and expanded context. Applying BNC corpus in real teaching is an encouraged teaching strategy to enrich students' vocabulary and also enable learners to produce accurate and natural sentences. The following Table 3 shows the first 10 nouns collocate with "play" in BNC corpus.

Table 3
The First 10 Nouns Collocate With "Play" in BNC

Rank	Nouns	Frequency
1	football	142
2	area	133
3	games	127
4	golf	93
5	tennis	73
6	rugby	58
7	cricket	51
8	cards	45
9	music	44
10	areas	41

Comparing with BNC, it is apparent that those cc3 collocation errors in the sentences above are mainly caused by literal translation and overgeneralization. From the Table 2, it is obvious that the noun collocates chosen by the native speakers to collocate with the verb "learn" in verb—noun pattern such as: learn English, learn thing, learn language, learn skills, etc. These collocations are listed for their high frequency used by native speakers. It is clear that Chinese students also use the verb "learn" frequently but it is only part of the meanings of "learn", which means equivalently with Chinese verb xue xi. Therefore, there exists in their writings a lot of verb—noun errors such as: learn the different cultures, learn the knowledge and theory, learn knowledge, learn more trends of economic growth and so on. In the CLEC, the verb "make" has almost only one meaning which is equivalent to the Chinese verb zuo, and for this reason Chinese students make errors by directly translate Chinese into English to make wrong collocations such as: make practice, make any social service. For overgeneralization errors, according to Ellis (1994) who pointed out that: overgeneralization errors arise when the learner creates a deviant structure on the basis of other structures in the target language. It generally involves the creation of one deviant structure in place of two target language structures. Chinese students are quite familiar with phrases: play football, play games, and play music. Students feel easy to generalize the rules and create some deviant collocations as play computer, play sports.

CONCLUSION

In the process of EFL learning, Chinese students will inevitably make some mistakes and teachers can get benefit from feedback of errors. The corpus-based approach is suggestive with its statistical reliability and the ability to present authentic and a large number of language data. The empirical study gives the meaningful suggestions which enable the author to propose pedagogical implications: based on CLEC, verb/noun collocation errors especially made by ST3 are the commonest and prominent problems when compared with others in "cc" categories in the process of Chinese students EFL learning and thus it should arouse the attention of teachers. By observation and analysis of the learner corpus, teachers can predict some errors which tend to be inappropriately used by students in their writings. Corpus of native speakers and all their findings are expected to help students to gain native-like command of English as target language. From statistical analysis to practical teaching implication, the present study is intended to shed new light on the collocation of pedagogy.

REFERENCES

BNC (The British National Corpus). Retrieved from http: // www.corpus. Byu.edu / bnc /

Ellis, R. (1994). *The study of second language acquisition* (p.59). Shanghai: Shanghai Foreign Language Education Press.

Firth, J. R. (1957a). A synopsis of linguistic theory, 1930-1955. *Studies in linguistic analysis*. Oxford: Basil Blackwell.

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

Gui, S. C., Yang, H. Z., & Yang, D. F. (2005). *Corpus-based analysis of Chinese learner English* (p.3). Shanghai: Shanghai Foreign Language Education Press.

Halliday, M. A. K. (1975). *Learning how to mean—Explorations in the development of Language* (p.75). London: Edward Arnold.

Laurence, A. (2005). *AntConc: Design and development of a freeware corpus analysis toolkit for the technical writing classroom* (pp.729-737). IEEE International Professional Communication Conference Proceedings.

Lou, B. (2001). Where did we go wrong? *A retrospective look at the design of the BNC*. Retrieved from http://users.ox.ac.uk/~lou/wip/silfitalk.html