Computer-aided contrast in Chinese L1 learners of English existential structures using Wordsmith Tools

Wenxiao Zhao[1, a)], Mingmei Du[2, b)]

[1]*School of Foreign Languages, University of Jinan, Jinan 250012, China.*
[2]*School of Foreign Languages, Shandong University of Finance and Economics, Jinan 250000, China.*

[a)]sfl_zhaowx@ujn.edu.cn
[b)]158307786@qq.com

Abstract. The present study has employed WordSmith Tools v3.9 as the retrieval program for the analysis and comparison of the complexity in English existential structures. CLEC and Brown are the corpus involved in the study with more than million words respectively, representing Chinese and native speakers' use of existential structures. WordList function has been served to generate keyword lists and the frequency of 11 types of existential structures while Concordance function, to capture contexts for the distribution of each type. Findings indicate that (1) Chinese L1 learners intend to overuse present tense and underuse past tense and perfect tense and (2) Chinese L1 learners showed lacks of flexibility in the expressions of different verbal tenses in using English existential.

Keywords: WordSmith Tools, English Learner Corpus, Chinese learners of English, existential structures, verbal tenses

INTRODUCTION

Existential structure, as a dynamic sentence structure which has its equivalent structure in most languages, appears frequently in Chinese learners' English compositions. Syntactically, there exist forms of past and future tenses, such as *there was* and *there will be,* as well as usage with alternative verbs in English, e.g. *there exist*. According to the previous studies of TG grammar, the standard form, its variants, as well as the conventional use of existential sentences are all acceptable forms by native speakers. According to notions of generativist, native speakers are able to use all the variants of existential sentences appropriately and intuitively. However, with regard to L2 learners, it is hard to manage all the variants of existential structures, but rather, rely on one or two variants more frequently than the other forms when speaking or writing English. As a result, the phenomena of overuse, underuse and misuse can be observed in L2 learners' writing production corpus.

In modern linguistics, corpus refers to a principled collection of natural texts with samples of a particular variety and uses of languages presented in a machine readable. This definition of corpus derives from a variety of definitions given by linguists in this field [1, 3]. Corpus linguistics can be defined as a linguistic methodology that is founded on the use of electronic collections of naturally occurring text or corpora [8]. Over the last three decades, the compilation and analysis of corpora stored in computerized databases have drawn great attention in the field of modern linguistics. There exist as many types of corpus as research topics in linguistics. Among those, learners' corpus is one of the most highlighted types, which have its advantages over other kinds of data in SLA research. According to Biber et al. [2], corpus-based approaches are general and concise as comprehensive studies of use cannot rely on intuition.

The present paper intends to examine differences between Chinese learners and native speakers in using existential structure by comparing their writing production using corpus and Wordsmith Tools, with the aim at providing new insights for teaching and learning of the mentioned grammatical item and a better understanding on how Chinese college students' writing differs from native speakers' varieties.

SUBJECTS, INSTRUMENT AND PROCEDURE

For this paper, we took CLEC corpus and Brown corpus as the research subjects. Each of the corpuses contains more than 1,000,000 words. Chinese Learner English Corpus (CLEC) consists of five sub corpora of writing contributed by Chinese students. The corpus is claimed to be reliable not only for the amount of sampled data, but also for the sampling process. The sampling proportion is balanced among learners at different levels from middle school to English major senior grades, and the samples are from diverse sources so that the corpus covers learners' written output widely enough [4]. Brown University Standard Corpus of Present-Day American English Corpus (Brown corpus) which compiled by Nelson Francis and Henry Kucera was available from 1964. It is a synchronic corpus of written English printed in the United States in the year 1961. It consisted of 500 samples, each of which is about 2,000 words of continues written English. Topics of Brown Corpus included news, religion, technique, and entertainment, etc. The total number of tokens is approximately 1,014,300 [3].

In order to retrieve all instances of English contrastive connectives in those two corpora, the present study employed WordSmith 3.9 as the retrieval program, which was designed by Mike Scott. It is a suite of lexical analysis tools operating under windows on text files stored on any drive. It can retrieve the search items specified by users from large amount of data, show the total number of its occurrence, and display all the occurrences of the search items in concordance lines of context, with the search items being highlighted on the screen. Three tools involved are WordList, Concordance and KeywordList.

We started with the WordList function to generate a list of ordered words that appeared in the target files. These can be used to compare the frequency of a word in different text files or across genres; within this tool, we compared two lists, lists for CLEC and Brown corpus. The words can be ordered either by frequency or by the starts of the word and also the list can be inverted. The present study used this tool to produce the list of the most frequent existential sentences in the Corpora Brown as well as in CLEC sub corpora of College students' compositions. By using the Concordance tool, we can obtain a number of examples of a specific word or phrase, in different contexts. The Concordance tool can generate Concordance lines from one or more target text chosen by users. In our study, we used Concordance function to get samples in lines for analysis. WordSmith can compare the clusters that appear in the target files with that appear in a reference corpus to generate a list of KeyWords. By using this function, we investigated the overused and underused 'there be' construction by Chinese ESL learners.

The whole study followed the sequence of a quantitative study first, and then a qualitative study. We collected all the constructions of *there be* and the variants, e.g. *there exist*, etc. in CLEC and Brown corpus by searching the keywords. We copied all the data in text files as the preparations of the process. With the help of WordSmith, we obtained the keyword list and the frequency of each form of existential structure. Then we used tables and figures to present results. The next step was to generate the Concordance lines of existential structure in the two corpora. Using these collected data, we investigated in what circumstances Chinese learners tend to underuse or overuse some forms of existential structure. And also, we explored factors that may attribute to misuses of existential sentences.

RESULTS AND DISCUSSION

Results: the overall frequencies and distribution

Table 1 demonstrates the frequency of each pattern of existential process in CLEC and Brown. We selected eleven types of existential process for our study, among which the type *there're* does not exist in Brown and the types like *there has been* and *there have been* cannot be found in CLEC. According to the statistics, the most frequently used forms in Brown are the present singular (*there is* 28.344%) and past singular (*there was* 29.370%) structures. Furthermore, balanced uses of the plural forms - 17.171% of *there are* versus 11.789% of *there were* - by native speakers can be observed.

In contrast, in the corpus CLEC, we surprisingly found more than half of the appearance of the structure constitute *there are* (54.984%). Furthermore, the total frequency of the present tense is over 86%, while the usages of other tenses such as past, perfect and future are less than 14% in total.

From **Figure 1**, it can also be observed that Chinese college students prefer to choose the simple present tense types than other varieties of existential structure. The percentage of 0% indicates that Chinese students seldom use perfect tense existential sentence, compared to native speakers. It is worth noticing the lack of some varieties of existential structure based on Chinese students' writing corpus compared to native speakers' corpus. Among the eleven types of existential sentence, the forms of *there had been, there have been* and *there has been* cannot be found in CLEC. Compared to data in Brown, the absence shows that Chinese ESL learners, to a certain extent, lack of flexibility in the expressions of different verbal tenses, especially in present perfect tense and the past perfect tense. **Figure 2** illustrates the different distribution of the tense in CLEC and Brown.

To conclude, from the data above, we can obtain the information that there exist overused types and underused types in Chinese learners' compositions. Chinese learners intend to overuse present tense and underuse past tense and perfect tense. Extra information presented in **Table 2** illustrates the frequency of the variants *there exist, -s, -ed* of existential structure in CLEC and Brown. However, due to the limited concordance lines in the two corpora, we cannot generalize the results of the comparison between corpora in this kind of variety of existential structures.

TABLE 1: The percentages of existential processes in different tense

Structures	**CLEC**		**Brown**	
	Lines	**Frequency (%)**	**Lines**	**Frequency (%)**
There is	405	32.556%	553	28.344%
There are	684	54.984%	335	17.171%
There's	11	0.884%	109	5.587%
There're	9	0.723%	0	0.000%
There was	52	4.180%	573	29.370%
There were	51	4.100%	230	11.789%
There has been	0	0.000%	23	1.179%
There have been	3	0.241%	20	1.025%
There had been	0	0.000%	39	1.999%
There will be	23	1.849%	35	1.794%
There would be	6	0.482%	34	1.743%
Total	1244		1951	

FIGURE 1: The comparison of the data between CLEC and Brown

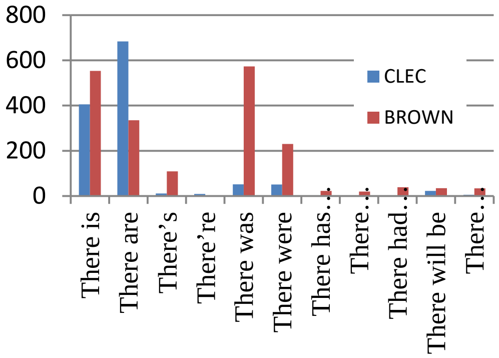
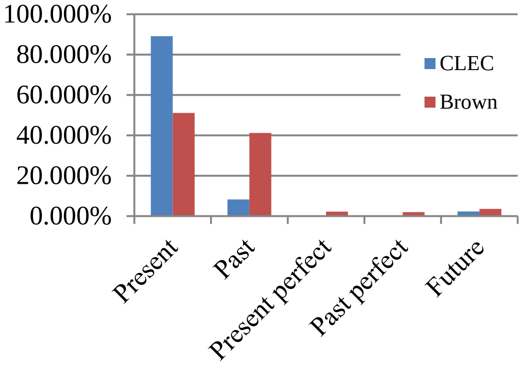


FIGURE 2: The different distribution of the tense used in CLEC and Brown.



Discussion: the causes of the overuse and the underuse

The causes of the overuse and the underuse of some varieties of existential structure by Chinese ESL learners lies in the following: First, in many grammar manuals compiled by Chinese scholars, the function of *there be* structure is highlighted like this: it can be used to emphasized the subject, the direct object of the infinitives, the prepositional object of the infinitive, the object of have-clause, the predicate of the original sentence and the attribute. More detailed, the exercises of these grammar manuals have been designed mainly with the transformation from a 'non-there-be' sentence to a 'there be' sentence, e.g., the sentence *China has several world-famous scientists* can be transformed to existential structure *There are several world-famous scientists in China* for grammar exercise. Apparently, this kind of exercises focuses mainly on the existents or the participant of existential process, but omitting the variants of the verb *be* which may lead to the overuse of the present tenses among Chinese learners.

To understand the problems that Chinese learners' underuse of the past tense and perfect tense, we need to know some elemental issues of L1 transfer as well as Chinese language morphology. Language transfer, as Odlin [6] defined, is the influence resulting from similarities and differences between the target language and other language that has been previously acquired. Newton and Kennedy [5] in their investigations of the effects of communication tasks on the grammatical relations marked by SL learners speculated that SL learners might be influence by the first language transfer. That is to say, when the L1 and L2 are different, the transfer is more likely to be negative, which can lead to transfer errors in the process of L2 acquisition. In the case of existential processes, as Chinese and English are different languages, L1 negative transfer is relatively inevitable in SLA.

TABLE 2: The distribution of the variants "there exist, -s, -ed" in different tense

Structures	CLEC		Brown	
	Lines	**Frequency (%)**	**Lines**	**Frequency (%)**
There exist	1	50%	2	28.571%
There exists	1	50%	2	28.571%
There existed	0	0	3	42.857%
Total	2		7	

The morphology of the Chinese language has neither inflections nor conjugations. Chinese speakers do not use past tense markers when they need to express activities in the past, but rather, using some particular words or phrase to replace the meaning of past tense and perfect tense. In this sense, it is possible that Chinese learners rely on their L1 transfer and thus overuse present tense and underuse past tense in their writing.

Krashen [7] suggested that learners can use the L1 to initiate utterances when they do not have sufficient knowledge of the target language. As a result, from an interlanguage point of view, we may also consider that the habit of Chinese learners in translating Chinese expression in mind literally into English without thinking about the target conventional expressions and the appropriate tense forms, or the lack of target language knowledge may also lead to the overuse and the underuse of certain varieties of existential sentences.

CONCLUSION

Through analyzing data from CELC and Brown, retrieved by WordSmith Tools, we have found that the existential sentences in native speakers' written material appear not only in present forms but also in past and perfect forms. Nevertheless, Chinese learners tend to rely mainly on simple present tense in their English compositions. Other tenses such as past tense and perfect tense are rarely used. The causes can be attributed to the English grammar manuals written by Chinese speakers, negative transfers from Chinese and learners' habit of translating Chinese into English literally. In this paper, we have focus only on the textual based corpus and eleven types out of 150 variants of existential structures, according to TG grammar. In the case of oral corpus, results could differ in a certain extent. And more comprehensive studies that cover all the possible subtypes and categorizations are needed to better illustrate this phenomenon in the future.

REFERENCES

1. A. Wilson, T, McEnery, (eds) Corpora in language education and research: A selection of papers from TALC94 UCREL, Lancaster University Lancaster, 1994.
2. D. Biber, S. Conrad, R. Reppen, Corpus linguistics: Investigating language structure and use. Cambridge University Press, Cambridge, 1998.
3. G. Kennedy, An introduction to corpus linguistics, Beijing: Foreign Language Teaching and Research Press, 2000.
4. H. Yang, S. Gui, D. Yang, A CLEC based analysis of Chinese English learners, Shanghai Foreign Language Education Press, Shanghai, 2005.
5. J. Newton, G. Kennedy, Effects of communication tasks on the grammatical relations marked by second language learners. System, (1996), 24, 309-322.
6. T. Odlin, Language Transfer. Cross-Linguistic Influence in Language Learning, CUP, Cambridge, 1989.
7. S. D. Krashen, Principles and practice in second language acquisition. Pergamon, Oxford, 1982.
8. S. Granger, C. Tribble, Exploiting Learner Corpus Data in the Classroom: form-focused instruction and data-driven learning, in: S. Granger (ed.), Working with Learner Language, Longman Harlow, 1998.