International Journal of Applied Linguistics & English Literature
ISSN 2200-3592 (Print), ISSN 2200-3452 (Online)
Vol. 1 No. 6; November 2012

Structural Analysis of Lexical Bundles in EFL English Majors' Theses of an Ordinary Normal University in China

Xixiang LOU

Faculty of Foreign Languages, Nanjing Normal University, Nanjing, China, 210097

Department of Foreign Languages and Literatures, Zhangzhou Normal University, Zhangzhou, China 363000

Tele: 13015611998 E-mail: Lou_xx@163.com

Received: 06-07- 2012 Accepted: 10-08- 2012 Published: 01-11- 2012

doi:10.7575/ijalel.v.1n.6p.142 URL: http://dx.doi.org/10.7575/ijalel.v.1n.6p.142

Abstract

A quantitative analysis has been made of 330 Chinese EFL learners' theses the distribution of the three- to eight-word lexical bundles in them and a comparison has been made of the percentages of the four-word lexical bundles of different structural categories in Chinese EFL learners' theses and the native English speakers' spoken or written academic language. It is found that the three-to eight-word lexical bundles in Chinese EFL learners' theses are on the decrease with the increase of the number of their component words. Chinese students' English language data share with native English speakers' spoken academic language data the 'personal pronoun + lexical verb phrase (+complement clause)' lexical bundles and the '(auxiliary +) active verb (+)' bundles, and also share with native English speakers' academic spoken language data the 'adverbial clause fragment' bundles, the 'noun phrase with other post-modifier fragment' bundles, the 'anticipatory it + VP/adjective P (+ complement clause)' bundles, the 'passive verb + PP fragment' bundles and the 'copula be + NP/adjective P' bundles. A further analysis shows that the EFL learners' English language in their theses is of more characteristics of written language and fewer characteristics of spoken language.

Keywords: Chinese EFL learners, theses; lexical bundles; structural categories; differences

1. Introduction

In recent decades, research interested in S/FL multi-word units has revived. Words are not just only considered as 'bricks' to fill the slots framed by grammatical rules, but they themselves have grammatical characteristics and they are the embodiment of grammatical rules. Cognitively, in a person's memory, words are not the only units of storage or access. Some multi-word units beyond individual words have also been widely recognized to be accessed and stored as a whole unit. They have become fixed or semi-fixed meaning-form pairings in language speakers' long-term process of contacting language materials. It is thought that a mastery of a large quantity of such units is facilitative to language users' fluent communication. Based on the degree of fixedness in their composition and the derivability of their meanings from their components, the multi-word units of different kinds form a continuum, on the one end of which are those which are put together according to the grammatical rules, and whose meanings can be obtained through understanding the meanings of their component words, and on the other end of which are those which are ungrammatical (at least based on modern grammatical rules they are so) and whose meanings are not the sum of the meanings of their component words and even some whose meanings seem to have nothing to do with their component words.

In such a field, terminology has always been problematic, for there is no generally agreed common term. Different terms are sometimes used to describe the identical or very similar kinds of units and at the same time, a single term may be used to denote very different phenomena (Moon, 1998: 2). Alison Wray（2002: 9）listed about 60 terms used for the same, similar or different kinds of multi-word units. In the past, the study of multi-word phenomena was mainly carried out from a perceptual point of view, the growth of computer corpora has facilitated the extraction of the multi-word units from a large amount of language data and has made it possible to make a large-scaled research on such phenomena in natural language. So it is more obvious to researchers, teachers and even students that multi-word units are ubiquitous in any language. *Lexical bundle* is a term firstly used by Bible (2000) to describe the multi-word units extracted from a language corpus with the computer. It was defined as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (ibid: 990). *Lexical bundle* is thought to be the extension of *collocation*, and the mutual combinations of words observed with the support of the computer technology. In recent years, it has attracted much attention and become a hot topic in language-dependent fields of research, e.g. cognitive linguistics, computational linguistics, psycholinguistics and applied linguistics. It is the same case with the field of SLA. "These bundles are familiar to writers and readers who regularly participate in a particular discourse, their very 'naturalness' signaling competent participation in a given community. Conversely, the absence of such clusters might reveal the lack of fluency of a novice or newcomer to that community".(Hyland, 2008)

2. Literature review on the empirical studies relevant to lexical bundles

A lot of research has been made on the distribution of lexical bundles in both L1 and L2 learners' productive language, or the differences or similarities between second language learners and native language speakers in their usage of lexical bundles. Biber et al.（2000: 987-1036) made a corpus-based study of the distribution of lexical bundles in English naive speakers' oral and written language. In their study, the lexical bundles in both spoken and written academic language were given a detailed description based on their structural categorization, and a comparison has also been made on the quantitative differences between the two registers in different categories of lexical bundles. It was found in the study that except for most categories of lexical bundles commonly used in spoken register and written register some categories of lexical bundles do not appear in spoken language or just occur at a low frequency, while conversely other categories of lexical bundles do not appear in written language or only occur at low frequency. Additionally, even among those appearing in both oral and written language, the coverage of the lexical bundles in two registers does not match completely. Such a result has made the differences between the two registers better known. Hyland(2008) found that "bundles are not only central to the creation of academic discourse, but that they offer an important means of differentiating written texts by discipline".

Lexical bundles are not only important in the fields of register and style research, they are also significant in L1 and L2 acquisition research, mainly involved in the description of language learners' proficiency development and the differences between L1 learners and L2 learners in language development. Cortes(2004) made a comparative analysis on the usage of lexical bundles in academic articles published by the professionals of history & biology and those in the theses of the learners of three grades of the corresponding specialties and found that the learners seldom used the lexical bundles usually used by the professionals. Shirato (2006, pp. 828-838) used the lexical bundle as one of his parameters for corpus analysis on Japanese EFL learners' English proficiency and English native speakers' language proficiency，showing that the two parties are significantly different in their usage of the two- to four-word bundles. Chen & Baker (2010) made a quantitative and qualitative comparative study on the lexical bundles in three corpora: one of the published academic texts and two of students' academic writing(one L1, the other L2). The results of the analysis showed that the published academic writing had the widest range of lexical bundles，while L2 students' writing had the smallest range. Furthermore, some high-frequency expressions in published texts were underused in both of the students' language corpora, while the L2 student writers overused certain expressions which native academics rarely used. Annelie & Erman (2012) made a quantitative analysis of the use of four-word English-language lexical bundles and also a qualitative analysis of the functions the lexical bundles serve in advanced learners' writing by L1 speakers of Swedish and in comparable native-speakers' writing, all writings being produced by undergraduate university students in the discipline of linguistics. It was found in the study that native speakers have a larger number of types of lexical bundles.

In China, research concerning lexical bundles has been carried out for several years. Much research focused on description of the distribution of lexical bundles in SL learners' productive language. Wang and Zhang (2006), after an analysis on the Chinese EFL learners' usage of lexical bundles in their argumentative compositions, drew the conclusion that Chinese EFL learners used fewer categories of lexical bundles and overused the 3-word bundles. Xu and Xu (2007) compared the Chinese EFL non-English major college students' usage of Discourse Management Chunks in their oral English speech with that of native speakers and found that Chinese EFL learners clung to literally translated chunks from Chinese and overuse the self-centered "I think" type of chunks which downgrades their discourse interaction to an ineffective and impolite mode. Ma (2009) extracted 191 high-frequency three-word lexical bundles from English native speakers' prose corpus, made a comparative study with Chinese EFL learners' usage of these three-word lexical bundles in their timed compositions and pointed out that the productive lexical bundles(including those overused ones) in the language learners' timed writing should be categorized as the SL learners' internalized and automated lexical bundles. Wei and Lei (2011) compared the use of the four-word lexical bundles in the doctoral dissertations by Chinese EFL learners with those in the published journal articles by professional writers in China and showed that the advanced learner writers used much more lexical bundles and much more different lexical bundles in their academic writing than professional authors. A structural analysis demonstrated that the advanced learners and the professional writers used similar amount of prepositional phrases, noun phrases, be + noun/ adjectival phrases and bundles of other structures. However, the learners used more passive structures and less anticipatory it structures of bundles than the professional writers. As for the functions of the bundles, the two groups of writers used a similar amount of research-oriented and text-oriented bundles. Nevertheless, the learners used less participant-oriented bundles than the professional writers. It was argued that the overuse of passive structures and the underuse of anticipatory it structures and participant-oriented bundles may be due to the learners' preference for the impersonality in their academic prose.

All these research has exposed many facts about the multi-word phenomena. However, the description is still not adequate. There are still a lot of unknown areas deserving further research. The usage of lexical bundles in the theses of the EFL university English major undergraduates in China is one of the blank areas.

For EFL university English major undergraduates in China, thesis-writing is the last and also most significant writing task. It is not only training but also a test. To call it training is due to that it is the first time for these learners to contact the relatively more formal academic writing. In some sense, thesis-writing is the enlightenment for their academic writing. And also theses writing will take students a long time to prepare, to write, to revise and to finalize. In fact, most universities have set the theses-writing course as selective or compulsory. In the process of writing, the students would be assigned a supervisor responsible for instruction on their theses writing. It is without any exaggeration to call such a process a process of training. To call theses-writing a test is due to the fact that the students have to pass the standards for the quality of their theses. Otherwise, they will fail to get their diplomas or others. These standards include requirements for language usage, structure, academic norm， etc. For the EFL learners, the fluent language usage is one of the most important requirements. In some sense, theses-writing is the final check on their language proficiency. So the theses written by the EFL university English major undergraduates are very important for both the students themselves and for the society. For the students, theses-writing has initiated their academic career. A high starting-point will be conductive to their future academic development. For the society, the students' good starting-points in academic writing will benefit the progress in the academic creativity of the whole society. Significant as the theses-writing is, the research on it is not adequate. What is the register difference between the EFL university English major undergraduates' theses and the academic language of native speakers? This is the topic of the study. The purpose of the study was to see the register characteristics of the EFL university English major undergraduates' theses through a comparative analysis of the usage of lexical bundles in three kinds of language data: the EFL University English major undergraduates' theses, native speakers' academic spoken language data and the native speakers' academic written language data.

The questions for the study are:

1) What is the quantity of the lexical bundles of different categories in the theses of the EFL University English major undergraduates' theses?

2) What structural categories of lexical bundles are there in EFL University English major undergraduates' theses?

3) What differences are there in the distribution of the lexical bundles of different categories in EFL University English major undergraduates' theses and native speakers' academic language?

3. Research design

3.1 The definition and categorization of the lexical bundles

Biber, et al(2000) defined the lexical bundles as "recurrent expressions, regardless of their idiomaticity and regardless of their structural status", "simply sequences of words forms that commonly go together in natural discourse"(990). They thought that the lexical bundles could be "extended collocations"(989). The lexical bundles are different from idioms. Idioms are usually of great fixedness in their structures and their meanings are hard to be derived from their component words. A lot of them are loaded with cultural or historical meanings. Additionally they are not necessarily so common in natural language. In contrast, the lexical bundles are multi-word units produced through the computer-based analysis of natural language data, and are the three-or-more-word units which are of statistically significant frequency of occurrence in language data. So "they are not fixed expressions, and it is not possible to substitute a single word for the sequence", and they are "much more common than idiom."(ibid)

Biber et al(2000) classified lexical bundles into 12 structural categories in their research on the lexical bundles used in English native speakers' oral academic language and written academic language：a）noun phrase with of-phrase fragment；b) noun phrase with other post-modifier fragment; c) prepositional phrase with embedded of-phrase fragment; d) other prepositional phrase fragment; e) anticipatory *it* + verb phrase/adjective phrase; f) passive verb +prepositional phrase fragment; g) copula *be* + noun phrase/adjective phrase; h) (verb phrase +) *that-clause* fragment; i) (verb/adjective +) to-clause fragment; j) adverbial clause fragment; k) pronoun/noun phrase + be (+…); l) Other expressions.

3.2 Language data

The language data for the study are 330 theses written by the students of 10 classes from an ordinary normal university in China. The length of the theses ranges from 5 000 words to 7 000 words. The topics of the theses include linguistics, literature, cross-cultural communication, translation, rhetoric and English for specific purposes. The main body part of the theses were used for analysis with the covers, abstracts, the list of contents，the list of references, acknowledgement, appendixes, and Chinese notes cut off. The size of the corpus is 1 673 529 tokens (1 544 872 tokens used for wordlist) or 37,908 types. The type/token ratio of the corpus is 2.30, and the standardized type/token ratio is 2.37.

3.3 The extraction of the lexical bundles

The extraction of the lexical bundles is based on two parameters: one is the frequency of the occurrence of the multi-word units in language data, usually called frequency cut-off point, counting the times of the occurrence of a lexical bundle in a million words(tokens); the other is the span of the multi-word units across different texts, i.e. a multi-word unit should not only appear in one text with a high frequency, which "guards against idiosyncratic uses by individual speakers or authors"(Biber 2006:134). However, the frequency cut-off is arbitrary. Biber et al (2000) set a ten-time-per-million-word frequency cut-off point for the four-word lexical bundles extraction, and a five-time-per-million-word frequency cut-off for the extraction of the five-word lexical bundles. And in another study (2006), he used a 40-time-per-million-word frequency cut-off. Cortes (2004) used a 20-time-per-million-word frequency cut-off. Biber and Barbieri (2007) used a conservative frequency cut-off— 40 times per million words. As for the requirement on the span to ensure the multi-word units are not from one speaker, Biber et al (2000) and Biber (2006) required that the high-frequency multi-word unit must be spread across at least five different texts to be counted as a lexical bundle. Cortes (2004) required that the word combinations should recur in five or more texts. In this study, the frequency cut-off will be decided on a 10-time-per-million-word frequency and the requirement of the span is to have coverage of at least five texts.

Wordsmith 4.0 was used to extract the three- to eight-word lexical bundles. The produced multi-word units together with the corresponding statistics on frequency and span were copied into an SPSS file and those failing to meet the requirements were deleted. The frequency of the three- to eight-word lexical bundles was respectively Calculated. Then the four-word lexical bundles were classified into different structural categories based on Biber et al's(2000) categorization of the oral and written academic language lexical bundles for a comparative analysis.

4. Result and Discussion

4.1 The distribution of three- to eight-word lexical bundles

In the study, there are 4592 three-word lexical bundles, 809 four-word lexical bundles, 145 five-word lexical bundles, 30 six-word lexical bundles, 10 seven-word lexical bundles and 5 eight-word lexical bundles extracted. The relationships between those adjacent categories of lexical bundles in quantity are as following: The three-word lexical bundles are 5.4 times as many as the four-word lexical bundles. The four-word lexical bundles are about 6 times as many as the five-word lexical bundles. The five-word lexical bundles are about 5 times as many as the six-word lexical bundles. The six-word lexical bundles are about 3 times as many as the seven-word lexical bundles. Lastly, the seven-word lexical bundles are about 2 times as many as the eight-word lexical bundles. Such a tendency is nearly the embodiment of the Zipf's law（http://www.nslij-genetics.org/wli/zipf/），and also the true reflection of the natural language. Such a tendency can be seen more clearly in the following bar chart.

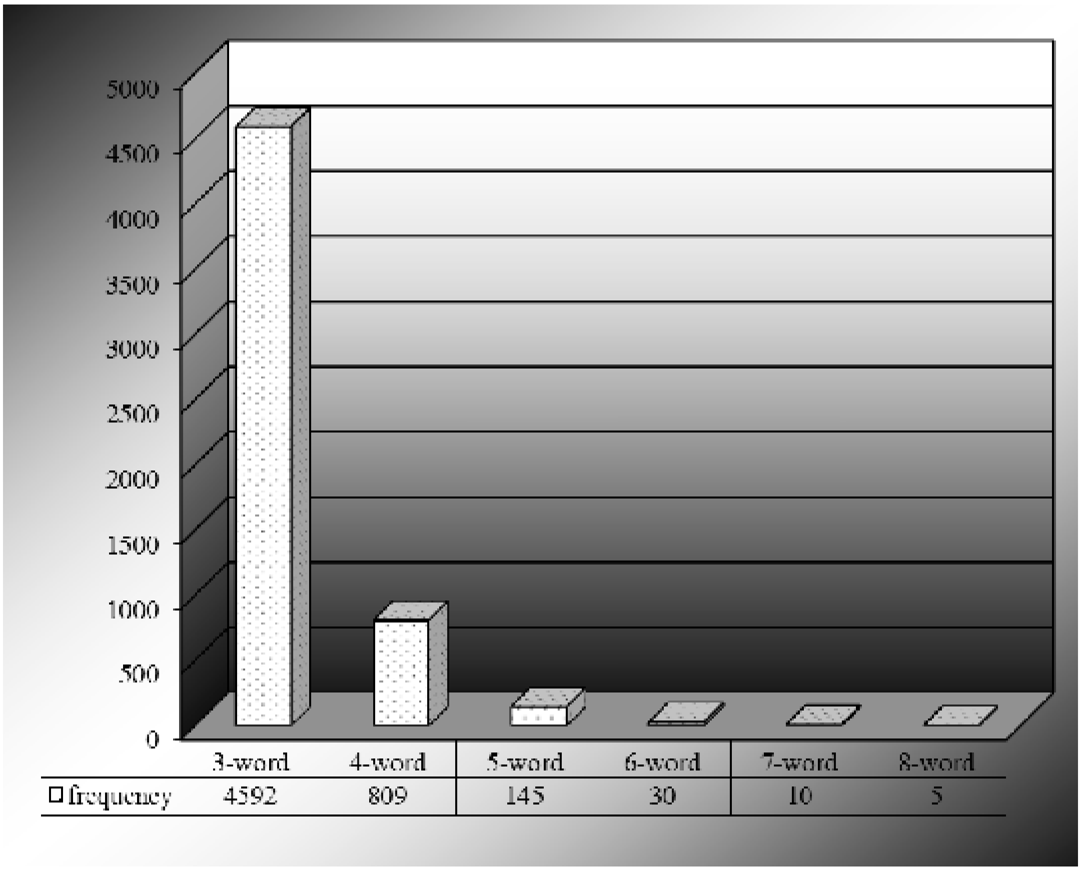


Figure 1. The quantities of the three- to eight-word lexical bundles

However, such a relation is different from that for native speakers' language data. Biber et al(2000) showed that in the English speakers' academic written language, the three-word lexical bundles were 10 times as many as the four-word lexical. The four-word lexical bundles were nearly 10 times as many as the five-word lexical bundles (the frequency cut-off was adjusted to '5 times per million words'.). There were no statistics for the other categories of lexical bundles.

In Biber et al's (2000) study, the top ten most common three-word lexical bundles whose frequency of occurrence were beyond 200 times per million words were (in descending order): ***in order to***, ***one of the***, *part of the*, *the number of*, *the presence of*, *the use of*, *the fact that*, *there is a*, *there is no*. In Chinese students' language data, the top 10 three-word lexical bundles whose frequency of occurrence was beyond 200 times per million words were: ***in order to, one of the***, *and so on*, *according to the*, *English and Chinese*, *the process of*, *Chinese and English*, *the development of*, *the meaning of*, *the target language*. Only the first two three-word lexical bundles overlap (in bold case). It seems that the Chinese students' lexical bundles are more topic-relevant, e.g. *English and Chinese, Chinese and English*, *the meaning of*, *the target language.*

In Biber et al's (2000) study, the most common four-word lexical bundles whose frequency of occurrence were beyond 100 times per million words were (in descending order)： *in the case of*, ***on the other hand***. In Chinese students' language data, the top 10 four-word lexical bundles whose frequency of occurrence was beyond 200 times per million words were: *at the same time*, *in the process of*, ***on the other hand***, *is one of the*, *one of the most*, *that is to say*, *an important role in*, *between Chinese and English*, *between English and Chinese*, *in the target language.*

4.2 The structural analysis and comparison of the four-word lexical bundles

Since it was the first time for the students to write theses, it was supposed that their language would be tinged with some oral language characteristics. The students' use of lexical bundles would more closely match those bundles found in conversation rather than those lexical bundles in academic prose (Cortes, 2002: 136). So in the study, the four-word lexical bundles extracted from the Chinese students' language data will be categorized and compared with Native speakers' four-word lexical bundles commonly used in their oral and written academic language which have been listed and categorized in Biber et al's (2000) research. In the process of classification, some unclassifiable but regular multi-word units are generalized into one category labeled as idiom and the rest 42 unclassifiable and non-sense multi-word units are deleted. The *idiom* refers to those high-frequency multi-word units which are independent of the main clause or those multi-word units labeled by Biber et al (2000) under the heading of idiom, e.g. *That is to say*, *More often than not* and *the same or similar*, etc. The statistics of the result of the treatment are summarized in table 1 and the number of different four-word lexical bundles across the major structural patterns in both English native speakers' spoken and written academic languages were listed in table 2:

Table 1. The categories, quantities and proportions of the lexical bundles used in Chinese students' theses

	Freq.	Per.
personal pronoun + lexical verb phrase (+ complement clause	28	3.5
pronoun/NP (+auxiliary) + copula *be* (+)	52	6.4
(auxiliary +) active verb (+)	44	5.4
(verb +) wh-clause fragment	1	.1
(and+)NP	77	9.5
quantifier expressions	1	.1
adverbial clause fragment	11	1.4
noun phrase with *of*-phrase fragment	144	17.8
noun phrase with other post-modifier fragment -	64	7.9
prepositional phrase with embedded *of*-phrase fragment	58	7.2
other prepositional phrase fragment	150	18.5
anticipatory it + VP/adjective P (+ complement clause) -	35	4.3
passive verb + PP fragment -	47	5.8
copula be + NP/adjective P -	51	6.3
(NP +) (verb +) that-clause fragment	11	1.4
(verb/adjective +) *to*-clause fragment	27	3.3
other expressions	5	.6
idiom	3	.4
Total	809	100.0

Note. Freq. = Frequency; Per. = percentage

Table 2. The number of different four-word lexical bundles across the major structural patterns in each register(from Biber et al., 2000:997)

Patterns more widely used in conversation	N.O.	N.W.
personal pronoun + lexical verb phrase (+complement clause)	187	—
pronoun/NP (+ auxiliary) + copula be (+)	33	5
(auxiliary +) active verb (+)	56	—
yes-no and wh-question fragment	49	—
(verb +) wh-clause fragment	17	—
(and +) NP	9	—
quantifier expressions	4	—
adverbial clause fragment	10	4
meaningless sound bundles	4	—

patterns more widely used in academic prose		
noun phrase with *of*-phrase fragment	16	69
noun phrase with other post-modifier fragment	—	15
prepositional phrase with embedded of-phrase fragment	6	56
other prepositional phrase fragment	7	35
anticipatory *it* + VP/adjective P (+ complement clause)	—	24
passive verb + PP fragment	—	16
copula be + NP/adjective P	—	11
(NP +) (verb +) *that*-clause fragment	3	13
patterns used in both registers		
(verb/adjective +) *to*-clause fragment	20	24
other expressions	3	5
Total	**424**	**277**

Note: N.O. =Native speakers' academic oral data; N.W. = Native speakers' academic written data

It can be seen from table 1 and table 2 that the lexical bundles used by Chinese EFL learners in their theses are not confined in the written language lexical bundles of native English speakers' language, their lexical bundles cover both written language and oral language in the language data by the native English speakers (excluding *Yes-no and wh-question fragment* and *Meaningless sound bundles*). In other words, it seems that judging from the usage of the lexical bundles, the Chinese EFL learners' English language in their theses is of both oral language characteristics and written language characteristics, and it is a kind of mixture of oral and written language. This is just the state of an interlanguage.

For a better comparison, both statistics of Chinese students' lexical bundles in their theses and the statistics of English native speakers' oral & written language lexical bundles were collapsed into one table and the frequencies of the lexical bundles are also converted into percentages. In the process, the statistics of the bundles labeled as 'idiom' were collapsed into that labeled as 'other expressions'.

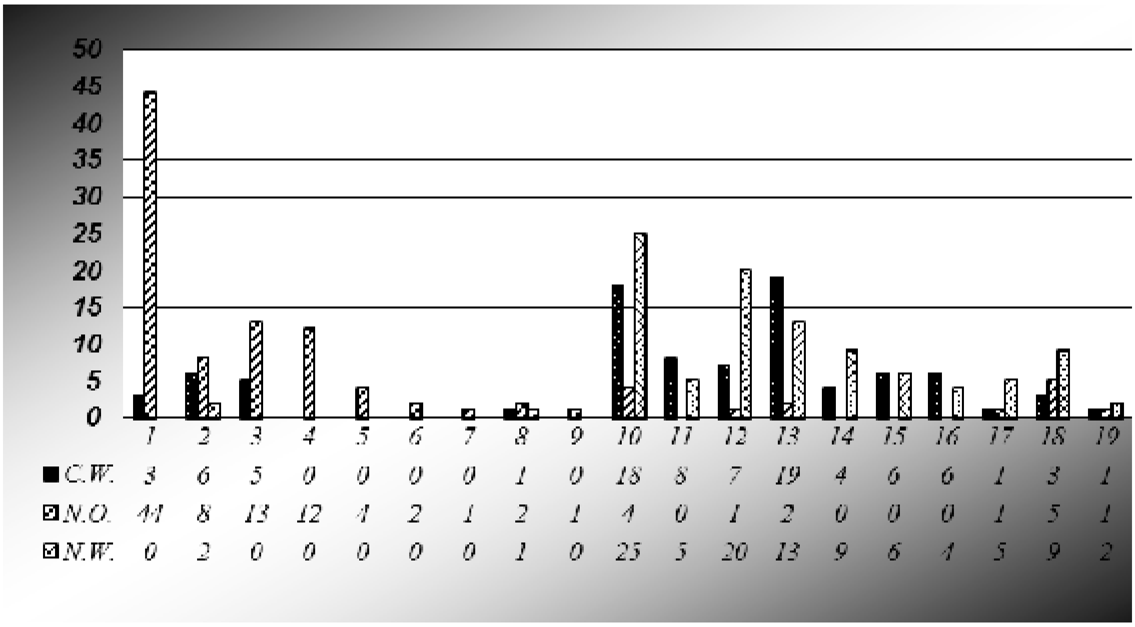
Table 3. The contrast between the percentage distribution of Chinese students' lexical bundles and that of the English native speakers' oral and written lexical bundles

patterns more widely used in conversation	N.O.	N.W.	C.W.
1.personal pronoun + lexical verb phrase (+complement clause)	44	0	3
2.pronoun/NP (+ auxiliary) + copula be (+)	8	2	6
3.(auxiliary +) active verb (+)	13	0	5
4.yes-no and wh-question fragment	12	0	0
5.(verb +) wh-clause fragment	4	0	0
6.(and +) NP	2	0	10
7.quantifier expressions	1	0	0
8.adverbial clause fragment	2	1	1
9.meaningress sound bundles	1	0	0
patterns more widely used in academic prose			
10.noun phrase with *of*-phrase fragment	4	25	18
11.noun phrase with other post-modifier fragment	0	5	8
12.prepositional phrase with embedded of-phrase fragment	1	20	7
13.other prepositional phrase fragment	2	13	19
14.anticipatory *it* + VP/adjective P (+ complement clause)	0	9	4

15.passive verb + PP fragment	0	6	6
16.copula be + NP/adjective P	0	4	6
17.(NP +) (verb +) *that*-clause fragment	1	5	1
patterns used in both registers			
18.(verb/adjective +) *to*-clause fragment	5	9	3
19.other expressions	1	2	1
Total percentage	100	100	100

Note：N.O. =Native speakers' academic oral data; N.W. = Native speakers' academic written data; C.W. = Chinese students' these data

In order to see the quantitative relationship among the language data in their lexical bundles, the statistics were made into a bar chart in figure 2, in which the quantities of the same category of lexical bundles in three language data were paralleled.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
■ C.W.	3	6	5	0	0	0	0	1	0	18	8	7	19	4	6	6	1	3	1
▨ N.O.	44	8	13	12	4	2	1	2	1	4	0	1	2	0	0	0	1	5	1
▨ N.W.	0	2	0	0	0	0	0	1	0	25	5	20	13	9	6	4	5	9	2

Note：N.O.=Native speakers' academic oral data; N.W.= Native speakers' academic written data; C.W.= Chinese students' these data

Figure2. The comparison among the quantities of the different-category lexical bundles in the three language data

Theoretically, there are two relation patterns involving Chinese students' language data for the distribution of a specific lexical bundle among the three kinds of language data. The first pattern is that a lexical bundle is shared by Chinese language data and either kind of native language data, i.e. oral or written language data. This means the Chinese language data are tinged with the characteristics of such kind of language registers, because it is probable that the lexical bundles of this kind profile the characteristics of this kind of register. There is another pattern, i.e. the three kinds of language data share a kind of lexical bundles. Probably, this kind of lexical bundles do not profile any specific language register. The first pattern deserves much notice, because it may reflect the oral or written language propensity of Chinese students' English language data.

It can be seen from figure 2 that the Chinese students' English language data share with the native English speakers' oral language data the following lexical bundles: **personal pronoun + lexical verb phrase (+complement clause)** and **(auxiliary +) active verb (+)**.

As for the 'personal pronoun + lexical verb phrase (+complement clause)'lexical bundles, in Chinese students' English language data, the lexical bundles beginning with 'we' lies on the top of the list of this kind of lexical bundles. Among 330 theses, 'we can find that' have covered 102 theses, 'we can see the' covers 29 theses. Such lexical bundles are usually used to extrapolate a conclusion from examples or data. They are of great subjectivity and usually not used in the writing of scientific reports (Liu, 2003; Tian & Duan, 2006). The propensity of the strong subjectivity in Chinese EFL students' writings has also been exposed by Wen's (2003) research, which showed that compared with American university students' compositions, Chinese advanced EFL learners used 2.84 times (90 times per million words) as many pronouns as the American peers in the ordinary argument writing. The theses belong to academic writing which is stricter in the amount of the usage of pronouns. The high percentage of pronouns in Chinese students' academic writing indicates that these students are still not completely familiar with the register characteristics.

In terms of the '**(auxiliary +) active verb (+)**'lexical bundles, the agents of their main verbs are usually human being, e.g. understand the meaning of，try their best to，try our best to，take part in the，should try their best，should pay more attention，should pay attention to，pay much attention to，pay more attention to，pay great attention to，pay attention to the，make use of the，make good use of，make full use of，learn from each other，know more about the，interact with each other，help the students to，have the ability to，have something to do，have a good command，have a better understanding，has something to do，guess the meaning of，does not want to，communicate with each other，attach great importance to，ask the students to，and ask them to，afraid of making mistakes, etc. Thus they no doubt reflect the characteristics of the oral language. Examples are not rare:

1) Therefore we should not only ***understand the meaning of*** the text, but also express it in the right way.

2) As we know, the better we ***understand the meaning of*** words, the better we will communicate smoothly.

3) If we want to know the definition of listening strategies, we had better ***understand the meaning of*** "strategy".

4) Only when have [siz] you known about the myths can you ***understand the meaning of*** these words exactly.

It can also be seen from figure 2 that Chinese students' English language data share with native English speakers' written language data the following lexical bundles: **adverbial clause fragment, noun phrase with other post-modifier fragment, anticipatory it + VP/adjective P (+ complement clause), passive verb + PP fragment** and **copula be + NP/adjective P**

As for the 'noun phrase with other post-modifier fragment' lexical bundles, in Chinese EFL learners' theses, there are the following lexical bundles: differences between Chinese and, relationship between language and, something to do with, differences between English and, an important part in, a good way to, the relationship between language, the relationship between the, much attention to the, the differences between Chinese, relationship between culture and, the differences between the, people from different cultures, nothing to do with, the fact that the, complex whole which includes, whole which includes knowledge, the best way to, the cultural differences between, great influence on the, the way in which, habits acquired by man, an effective way to, the ways in which, a great influence on, differences between the two, the relationship between culture, information usually paid for, the differences between English, ideas by identified sponsors, a great impact on, carrier of culture and, the distance between the, people from different cultural, English teaching in china, a person who is, cultural differences between Chinese, relationship between teachers and, similarities between Chinese and, people from English speaking, differences between the Chinese. Among these lexical bundles, the way in which, the ways in which, the fact that the, the relationship between the, the differences between the, an important part in were listed on the top of the list of native speakers' written language lexical bundles, among which, 'The way in which' , 'the fact that the' were used over 40 times per million words, and 'the ways in which' , 'the relationship between the' have been used over 20 times per million words.

The '**anticipatory it + VP/adjective P (+ complement clause)**' lexica bundles are very commonly used in written English language to convey a range of epistemic, evaluative, and attitudinal meanings (Jalali et al, 2009). Most extraposed complement clauses beginning with anticipatory 'it' can also reflect the speaker or writer's assessment (Hewings and Hewings, 2002, qtd. in Jalali et al, 2009). This kind of lexical bundles are not so frequently used in native English speakers' oral language data, but they are very frequently used in both native English speakers' written language and Chinese EFL learners' theses. In Chinese EFL learners theses, there are the following lexical bundles: it is necessary to, it is necessary for, it is obvious that, it is difficult to, it is said that, it is important to, it is very important, it is impossible to, it is easy to, it is believed that, it is better to, it is not only, it is clear that, it is hard to, it is true that, it is one of, it is a good, it is well known, it is not the, it is hard for, it is the most, it is impossible for, it can be seen, it is difficult for, it is not easy, it is the same, it is very difficult, it does not mean, it is also the, it is found that, it is a very, it comes to the, it is of great, it means that the, it is important for, it is not a, it is easy for, it is known that. Among these lexical bundles, ‘it is impossible to’, ‘it is necessary to’ were used over 40 times per million words, and ‘it is important to’, ‘it is clear that’, ‘it is difficult to’, ‘it can be seen’ were used over 20 times per million words and ‘it is easy to’, ‘it is true that’ have been used over 10 times per million words.

As for the ‘**passive verb + PP fragment**’ bundles, there are the following lexical bundles used over 10 times per million words in Chinese EFL learners’ theses: can be divided into, can be used to, is closely related to, can be found in, closely related to the, is based on the, are closely related to, be divided into two, is considered to be, can be regarded as, can be defined as, can be classified into, is regarded as a, used to refer to, be found in the, influenced and shaped by, considered to be the, be regarded as a, can be used in, is well known that, be taken into consideration, is considered as a, is made up of, is considered as the, is defined as the, related to each other, can be seen in, is often used to, is regarded as the, is widely used in, is often used in, is used as a, be regarded as the, can be used as, be divided into three, pointed out that the, should be taken into, the above we can, be used as a, be considered as a, should be translated into. Among these lexical bundles, ‘is based on the’ has been used over 20 times per million words, ‘can be found in’, ‘be found in the’, ‘be used as a’ have been used over 10 times per million words in English native speakers’ written language data.

In terms of the ‘**copula be + NP/adjective P**’ bundles, there are the following lexical bundles used over 10 times per million words in Chinese EFL learners’ theses: is the most important, is not only a, are a lot of, is a part of, is the symbol of, something to do with, is not only the, is a good way, is very important to, is the process of, is a process of, is of great importance, is quite different from, is the carrier of, is more important than, is very important for, is a very important, be aware of the, is well known that, is an important part, are two kinds of, was one of the, is no doubt that, is very important in, is the use of, is a reflection of, is the same as, is a symbol of, be one of the, is not easy to, is the basis of, is a system of, is the result of, is necessary for us, is an island country, is the carrier and, is similar to the, is one kind of, are more likely to, is a type of, is the core of, is known to all, is the key to, is a sign of, is a figure of. Among these lexical bundles, ‘is one of the’ has been used over 20 times per million words in native speakers’ written language data, and ‘is the result of’ has been used over 10 times per million words.

From the above elaboration, it does seem that Chinese EFL learners’ theses are of both spoken and written English language Characteristics seen from the perspective of the lexical bundles. However, it is not completely right to conclude that Chinese EFL learners’ English language in their theses is a balanced mixture of oral and written language. In fact, it is of more written English characteristics than spoken English characteristics.

Seen from a macro perspective, among the lexical bundles more widely used in English native speakers’ academic conversation, there are more that can not be found in Chinese EFL learners’ these: ‘yes-no and wh-question fragment’, ‘(verb +) *wh*-clause fragment’, ‘(and +) NP’, and ‘quantifier expressions’. In contrast, almost all categories of lexical bundles significant in native English speakers’ academic written language can also be found outstanding in Chinese EFL learners’ these.

Seen from a micro perspective, even in those lexical bundles shared by Chinese EFL learners’ theses and native English speakers’ spoken academic language data, there is still a gap between Chinese EFL learners and the native English speakers. In native English speakers’ oral data（Biber et al, 2000: 1002-1003), the frequently used four-word bundles include mainly the expressions with ‘I’: a) **expressions with *I* + *know*:** I don't know what+, well I don't know”, “I don't know how+, I don’t know if+, I don't know whether+, I don't know why, oh I don't know, 'but I don't know, I don't know where+, I didn't know that, I didn't know what, and I don't know, so I don't know, I don't know about, I don't know I, I don't know who, I don't really know, yeah I know but, I know what you+, I mean I know; b) **Expressions with *I* + *think***: +I don't think so, but I don't think, I don't think he, '-I don't think I+, I don't think it+, +I don't think it's, I don't think you+, +no I don't think+, well I don't think, I don't think she, I don't think that, I don't think they, I don't think we; I thought it was+, I thought 1 would, I thought that was, I thought you were, I would have thought, and I thought oh, and I thought well, so I thought well, I thought he was, I thought they were, I thought to myself; I thought you said; I think it was, I think I might, I think I would, I think it's a, I think it is, I think you should, I am thinking of; c) **Expressions with *I* + *want*:** I don't want to+, I didn't want to, but I don't want+, I don't want it, no I don't want+; +I want to do, I want to get, 'I want to go+, +I want to see, and I want to, I just want to, I want to be, I want to know, I want you to; d) **Expressions with *I* + *said/tell***: +I said to him, and I said to+, so I said well, +I said to her, I said well I, and I said I, and I said oh, and I said well, so I said to, I said I don't, I said 1 would; I tell your what+, I'll tell you what. Expressions with I + like: I would like to+; I don't like it, I don't like that, I don't like the, I don't like them; e) **Expressions with *I* + *mean*:** I mean I don't+, but I mean I, yeah but I mean, I mean if you, I mean it's not, I mean it was, I mean you know; f) **Expressions with *I* + modal/semi-modal verb:** I was going to+, I'm not going to, I'm going to do, I'm going to get, I'm going to go, I'm going to have+, I'm going to get, I'm going to put, I'm just going to, well I'm going to; I would have to, I would love to, I would rather have, I would have been; I'll give you a, I'll go and get, I'll have a look, I'll have to go, I'll have to get; I shall have to; I don't have to, I had to go, I have to go; I've got to go, I've got to do, I've got to get; I can't be bothered, I can't do it, I can't remember what, I couldn't believe it; I might as well; I used to go; g) **Other expressions with *I* + verb phrase:** I haven't got a, I haven't got any, I see what yon+, I went to the, I was talking to, I was trying to; h) **Expressions with *you* + *know*:** +you know what I f , you don't know what, you know when I, you know when you, you know where the, you know I mean; i) **Expressions with *you* + *want*** "+you don't want to, +you want me to+, +you want to go, "you want to do, you want to be, +you want to come, you want to get, you want to see; j) **Expressions with *you* + modal/semi-modal verb:** you don't have to, you have to do, and you have to, you have to go, you have to have, you have to pay, well you'll have to, you're going to have, you're not going to, you were going to, you're not going to, if you're going to, you're going to get; yon can have a, and then you can, you can do it, you can get a, you can have it; you've got to be, you've got to do, you've got to go, you've got to have; you have to be; you might as well; you'll be able to, you're not supposed to, you don't need to; k) **Other expressions with *you* + verb phrase**: you see want I+, +you think about it, you haven't got a; l) **Expressions with *he/she* + said**: he said to me, and she said oh, she said to me; m) **Other pronoun + verb phrase expressions:** he was going to, he's not going to, she was going to; 'we're going to have, we were going to, we're not going to, we used to have; '+us have a look+; they're not going to, they were going to, they don't want to; it's not going to, it was going to. While in Chinese EFL learners' theses, there are not so many variant categories of such lexical bundles and not so great an amount of such lexical bundles. They are mainly expressions with 'we': *we can see that*, *we can see the*, *as we know the*, *we can find the*, and *we know that the.* There is only one lexical bundle with *they: they do not know*, and one lexical bundle with I: *I would like to*. Comparatively speaking, the expressions with 'we' have a very low degree of subjectivity when out of concrete context, especially in written language. Usually 'we' is used to refer to anyone reading the written academic information. In such a case, 'we' has got some degree of objectivity.

5. Conclusion

From the above analysis it can be concluded that, the three- to eight-word lexical bundles in Chinese EFL learners' theses are on the decrease with the increase of the number of their component words. As far as the four-word lexical bundles are concerned, Chinese EFL learners' language in their theses share some categories of lexical bundles with native English speakers' oral academic language and share some other categories of lexical bundles with native English speakers' written academic language. However it is not a balanced combination of oral English language characteristics and written English language characteristics, but one of more written language characteristics and fewer oral language characteristics. Chinese EFL learners' written language has both the characteristics of oral English and that of written English, which has been shown through relevant word studies, e.g. Wen and Ding (2004), Zhang (2007). Theses-writing is different from daily writing, because it should be of more academic characteristics.

The potential causes for such characteristics of Chinese EFL learners' English language in their theses may be: firstly, recent years' emphasis on oral English has weakened the cultivation of students' writing abilities. Much of the students' attention has been distracted away from reading, which has delayed the formation of students' writing abilities; secondly, it is the first time for the undergraduate students to try academic writing and most of them do not make empirical reports but just do literature -review writing. So their knowledge of academic norm and register is still not adequate; thirdly, teachers have not attracted students' attention to the register characteristics of academic writing, which leads to students' weak awareness of the register characteristics of academic language.

Based on the above analysis, in order to cultivate EFL learners' ability in academic writing, in language teaching, students need to be allowed enough time for reading academic writings. Teachers can purposely attract students' attention to the characteristics of academic language and to the differences between oral and written language. Students should be reminded of not implanting oral language characteristics into academic writing.

References

Annelie, & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*. *31*(2), 81-92.
Biber, D. (2006). Lexical Bundles in University Teaching and Textbooks *University Language: A Corpus-based Study of Spoken and Written Register*. Amsterdam: John Benjamins Publishing Company.
Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.
Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2000). *Longman Grammar of Spoken and Written English*. Beijing: Foreign Language Teaching and Research Press.
Chen, Y.-h., & Baker, P. (2010). Lexical Bundles in L1 and L2 Academic Writing. *Language Learning & Technology, 14*(2), 30-49. Retrieved from http://llt.msu.edu/vol14num2/chenbaker.pdf
Cortes, V. (2002). Lexical bundles in Freshman composition. In R. Reppen, S. M. Fitzmaurics & D. Biber (Eds.), *Using Corpora to Explore Linguistics Variation*. Philadelphia: John Benjamins Publishing Company.
Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23 (4),* 397–423.
Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27* 4-21.
Jalali, H., Rasekh, A. E., & Rizi, M. T. (2009). Anticipatory 'it' lexical bundles: A comparative study of student and published writing in applied linguistics. *Iranian Journal of Language Studies (IJLS) 3*(2), 177-194.
Shirato, J. (2006). *Using Learner Corpora to Teach Authentic English.* Paper presented at the JALT 2005, Tokyo.
Wei, Y., & Lei, L. (2011). Lexical Bundles in the Academic Writing of Advanced Chinese EFL Learners[J]. *RELC Journal, 42*(2), 155–166.
Liu X.M. (Ed.). (2003). *Standard Formals for English Research Papers*. Beijing: Higher Education Press.
Ma G.H. (2009). Lexical bundles in L2 timed writing of English majors. Foreign Languge Teaching and Research*, 41*(1), 54-60.
Moon, R. (1998). *Fixed Expressions and Idioms in English: A Corpus-Based Approach*. Oxford: Clarendon Press.
Tian G.S., & Duan X.Y. (Eds.). (2006). *Writing Graduation Thesis*. Beijing: Beijing Institute of Technology Press.
Wang L.F., & Zhang Y. (2006). A Corpus-based Study on Chunks in English Argumentative Writing of Chinese EFL Learners. *Computer-assisted Foreign Language Education,* (4), 36-41.
Wen Q.F.& Ding Y.R.(2004). A Corpus-based Analysis of Frequency Adverbs Used by Chinese English Majors. *Mordern Foreign Language(Quarterly),27*(2):150-156.
Wen Q.F.,Ding Y.R. & Wang W.Y. (2003). Features of Oral Style in English Compositions of Advanced Chinese EFL Learners: An Exploratory Study by Contrastive Learner Corpus Analysis. *Foreign Language Teaching and Research(Bimonthly),35*(4):268-274.
Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.
Xu J.J., & Xu Z.R. (2007). Discourse Management Chunks in Chinese college learners' English speech: A spoken corpus-based study. *Foreign Language Teaching and Research(Quarterly), 39*(6), 437-442.
Zhang P. (2007). A Corpus-based Contrastive Analysis on Lexical Complexity of Chinese and International EFL Learners, *Foreign Language of China*, *4(3)*: 54-59.

.