English Language Teaching; Vol. 13, No. 1; 2020
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Lexical Bundles in Thesis Abstracts by L1 Chinese Learners of English and U.S. Students

Meng Lyu[1] & Roger W. Gee[2]

[1] School of Foreign Studies, North China University of Water Resources and Electric Power, Zhengzhou, China

[2] School of Education, Holy Family University, Philadelphia, USA

Correspondence: Meng Lyu, School of Foreign Studies, North China University of Water Resources and Electric Power, Zhengzhou, Henan Province, China, 450046.

Received: November 26, 2019 Accepted: December 21, 2019 Online Published: December 23, 2019

doi: 10.5539/elt.v13n1p141 URL: https://doi.org/10.5539/elt.v13n1p141

Abstract

The general question this research investigates concerns the difference between the use of lexical bundles in a corpus of abstracts for theses in the liberal arts written by Chinese undergraduate students and a corpus of abstracts written by American master's degree students. The undergraduate abstracts were first written in Chinese and then translated in to English at a medium-sized university in China. The master's degree theses abstracts were downloaded from an online database. It was found that there were differences in the types and tokens of lexical bundles in the two corpora with few shared bundles. There were fewer differences in the structural characteristics and in the functions of the lexical bundles found in the two corpora. Specific pedagogical recommendations are made, as well as implications regarding methodology in future research.

Keywords: lexical bundles, abstracts, L1 Chinese, learner corpus, academic writing

1. Introduction

It has been found that language is, to a great extent, formulaic. Conrad and Biber (2005) conclude that multi-word sequences are used with a high frequency by native speakers and writers. Both Hyland (2008b) and Nation and Webb (2011) noted that multi-word sequence plays an important role in fluent language production and successful language learning. As might be expected due to the number of studies of formulaic language, there are a number of terms used to refer to multi-word sequence, including clusters, formulaic sequence, lexical bundles, recurrent combinations, chunks, and n-gram. The term "lexical bundles" used by Biber, Johansson, Leech, and Finegan (1999), will be used in this article, though not to differentiate lexical bundles from other terms.

Over the last 20 years, there has been a large body of research exploring the use of lexical bundles through corpus analysis. Most of these studies focus on lexical bundles in academic writings and research journals. Some compare the differences in use of lexical bundles between student and expert writers, native and non-native scholars, between L1-English and L2-English student writers. Quite a few studies have compared the use of lexical bundles in abstracts of academic writing by English L1 and particular language groups writing in English, such as Chinese. Therefore, the present study is to compare the use of lexical bundles in English- language academic abstracts by L1-Chinese and native-speakers and is expected to give insights to non native instructors to help L2 students use the specific lexical bundles and produce more coherent and more professional academic abstracts.

1.1 Review of Literature

Biber and his colleagues ((Biber et al., 1999; Conrad & Biber, 2005; Biber & Barbieri, 2007) have conducted a series of studies to explore lexical bundles in speech and writing. They found that speakers and writers use lexical bundles regularly and different registers are characteristic of different sets of lexical bundles to build the particular discourse. Hyland (2008b), Hyland (2012) and Durrant (2017) explored the use of lexical bundles in different disciplines of academic writing, which showed the considerable disciplinary variation of lexical bundles in academic writing. These studies suggest that learning the most frequent lexical bundles in different genres help learners gaining communicative competence in building academic discourse.

A number of studies have investigated lexical bundles in second language (L2) students' academic writing and

made comparison with first language (L1) student writing as well as with more proficient writers including published research articles by both native and non-native speakers of English. Hyland (2008a) compared research articles, PhD Dissertations and MA/MSc theses from different disciplines and found less proficient students in constructing their texts rely greatly on lexical bundles. Güngör and Uysal (2016) compared research articles written by native and non-native scholars, which found non-native scholars used a large number of and more varied lexical bundles than the native scholars did. Pan, Reepen, and Biber (2016) conducted the study on lexical bundles of L1 versus L2 English academic professionals in research journals. Their study also found that there are a greater number of bundles in the L2-English corpus. Chen and Baker (2010) compared native expert writing and L1 English students and L1 Chinese students of L2 English. They found that the number of lexical bundles increases with advancing writing proficiency. All these studies have made structural and functional comparisons of lexical bundles. Because of different corpora involving different disciplines and different writers, the conclusions vary.

While many studies focus on lexical bundles across a wide range of academic disciplines, several researchers have explored the lexical bundles in single domain or specific discourse settings. Pan et al. (2016) carried out a comparison of Telecommunications research articles. Shin (2018) compared lexical bundles in argumentative essays written on the same topic by native and non-native writers. Alasmary(2019) investigated aspects of lexical bundles in Mathematics texts and presented a list of mathematics-oriented lexical bundles . These studies demonstrate that lexical bundles are discipline-specific and genre-specific. The present study focuses on one specific academic genre, abstracts.

1.2 Overview of the Present Study

The present study examines abstracts for theses in the liberal arts written by undergraduate students in China and master's degree students in the United States. Graduation thesis writing is an important part of an undergraduate education in all universities in China. English majors write their theses in English, while in many universities non-English majors write their theses in Chinese and translate the abstract into English.

The general question this research investigates concerns the difference between the use of lexical bundles in a corpus of abstracts written by Chinese undergraduate students and a corpus of abstracts written by American master's degree students. More specific questions are:

1) What are the differences in the types and tokens of lexical bundles in the two corpora?

2) What are the differences in the structural characteristics of the types and tokens of lexical bundles in the two corpora?

3) What are the differences in the functions of the types and tokens of lexical bundles in the two corpora?

The purpose of investigating these questions is to provide guidelines about how the abstracts written by the Chinese non-English major students can sound more natural.

The paper is organized as follows: Section 2 introduces the construction of two matched corpora, CHTAs and USTAs, and identifies lexical bundles used as the basis for the present study. Section 3 presents the major findings of the study and analyzes lexical bundles structurally and functionally. Section 4 discusses the results of this study and compares them to other studies of L1 and L2 writers of English. Finally Section 5 discusses conclusions and implications.

2. Method

2.1 Corpus Construction

Two matched corpora were constructed. The first was a corpus of undergraduate thesis abstracts (CHTAs) written by students as a graduation requirement from a middle-sized university in northern China. The abstracts were written by students who graduated in 2016. The second corpus was constructed from the abstracts of master's thesis (USTAs) written by students in a number of universities in the United States. These abstracts were written in the time period from 2013-2017 and retrieved from a database, ProQuest Dissertations and Theses in the spring of 2017. The two corpora are comparable in the sense that they have an equal number of abstracts in the content areas of linguistics, literature, culture, pedagogy, and translation. However, the two corpora do contain a different number of words. Details about number of words in each corpus can be seen in Table 1.

Table 1. Corpora of Chinese and U.S. –based Thesis Abstracts (Calculated by Antconc ver.3.4.4)

	CHTAs		*USTAs*	
Academic divisions	*No. of abstracts*	*Words*	*No. of abstracts*	*Words*
Culture	88	27,771	88	17,384
Linguistics	50	15,741	50	11,322
Literature	167	52,777	167	34,338
Pedagogy	10	2,754	10	2,577
Translation	2	736	2	399
Total	317	98,297	317	65,021
Mean Words		310		205

Overall, Chinese thesis abstracts in the corpus are longer than U.S.-based abstracts. That may be because the Chinese students were required to write abstracts with no less than 300 words. On the other hand, the US-based abstracts were selected randomly with no minimum words requirement. Each US abstract has 205 average words. Since the size of the two sub-corpora is different, the standardized frequency was used to compare the bundles across sub-corpora of different sizes.

2.2 Identification of Lexical Bundles

After constructing the two corpora, lexical bundles were identified, following conventional procedures. However, as the identification of the lexical bundles required several steps, we go into some details to describe the steps and some of the decisions that were made as work progressed. We believe that the transparency achieved by a clear explication of what was done will make the results more meaningful.

The first step in this study was to extract all 4-gram bundles from the two sub-corpora using AntConc (version 3.4.4). Following Simpson and Ellis (2010), 2-word bundles were not included as 2-word sequences are highly frequent and are often subsumed in 3- or 4-word phrases. It appeared that the 5-grams contained overlaps with the 4-grams, as Biber et al. (2004) also observed. In addition, as Adel and Erman (2012: 84) pointed out that "three-word bundles are often subsumed in four-word bundles". Simpson and Ellis point out, there is a need to keep the data at a manageable size. and Baker (2010:32) noted that "the number of 4-word bundles is often within a manageable size". According to Hyland (2008b: 8), 4-word bundles "offer a clearer range of structures and functions than 3-word bundles". It was thus decided to extract 4-grams, which was in line with previous research that analyzed only 4-grams.

The next step was to exclude those n-grams that were not frequent. Previous research use cut-off ranges between 10 and 40 instances per million words. A conservative approach to identify lexical bundles is a relatively high frequency cut-off point. Since both the Chinese and the American corpora were rather smaller, we used a more conservative figure of 60 times per million, to avoid a situation where a few bundles were identified on the basis of a chance occurrence. Thus, a raw frequency cut-off for considering initial lexical bundles was set at a minimum of 6 times and 4 times for the CHTAs corpus and USTAs corpus, respectively. These cut-off points were normalized per 1000 words as shown in Table 2, where it is seen that a normalized frequency per 1000 words is equal to 60 times per million.

Table 2. Raw and Corresponding Normalized Frequency Thresholds Adopted

Corpus	*Corpus size*	*raw frequency*	*normalized frequency per 1000 words*
	1,000,000	60	.06
CHTAs	98,297*	4	.06
USTAs	65,021*	6	.06

After eliminating those n-grams that did not meet the frequency standard, we looked at the distribution of the remaining bundles. Biber et al. (2004) state that an n-gram must occur in at least 5 different text to be counted as a lexical bundle, though they note that this distributional requirement has little practical effect as noise bundles are widely distributed. Nevertheless, to avoid against distinctive overuse by individual writers, it was decided that lexical bundles had to appear in at least two different academic divisions to be included into the final

analysis.

The qualitative identification involved a manual analysis of the 4-grams. We began with deleting 4-grams containing proper nouns including names of people, dialects, languages, cities, and countries; titles of television shows, books, and websites; festival names; and historical events such as dynasties. Four-grams with punctuation such as commas, periods, quotation marks, and apostrophes were also deleted to avoid 4-grams formed by spanning two clauses or sentences. Finally, 4-grams with numbers and abbreviations were deleted.

Overlap among lexical bundles is an issue when counting. Overlap occurs when "lexical bundles of fixed length ... are fragments of longer chunks of text" (Grabowski & Juknevičienė, 2016: 58). That is, two 4-grams might not be mutually exclusive as they would both contain the same 3-gram. For example, *the ways in which* occurs 24 times and *ways in which the* occurs 4 times, but the occurrences of *ways in which the* subsume those of *the ways in which*. The problem, if no correction is made for overlap, is that "Overlapping word sequences could inflate the results of quantitative analysis" (Chen & Baker, 2010: 33). Chen (2009: 67) goes into some details about how the problem of overlap may be resolved, but concludes that "it can be seen that the system outlined here is both methodologically and perceptually complex. It also has to acknowledge that this system does not fully resolve the problem of over-representation or under-representation in determining lexical bundles".

Rather than attempt to use the system proposed by Chen (2009), we found that most cases of overlap could be resolved by visual examination of color-coded bundles. Cases of possible overlap were identified in the lists of cleaned 4-grams. Then concordance lines for the common 3-gram were retrieved with AntConc and were examined by the two researchers together.

Each instance of possible overlap was color coded. Figure 1 is an example, using bold-face rather than color-coding to accommodate a lack of color. The 3-gram *thesis focuses on* is used in two 4-grams, *this thesis focuses on* and *thesis focuses on the.* Only the more frequent 4-gram, *this thesis focuses on,* which occurred seven times, was retained. The other 4-gram, *thesis focuses on the,* which occurred four times, was not counted to avoid an overcount. Had it been counted, 11 lexical bundles would have been reported when there were only seven. Another example was with the 3-gram *the structure of.* It appeared 18 times. Sixteen occurrences were as part of the 4-gram, *the structure of the.* The 3-gram *the structure of* also appeared 10 times as part of the 4-gram *from the structure of.* However, nine of these overlapped the 4-gram *the structure of the*, and had already been counted. Only one instances of *from the structure of* had not been counted, and this one time did not meet the frequency criteria. Thus, the 4-gram *from the structure of* was not counted as a separate lexical bundle. The number of lexical bundles remaining after the removal of overlapping bundles is presented in Table 3.

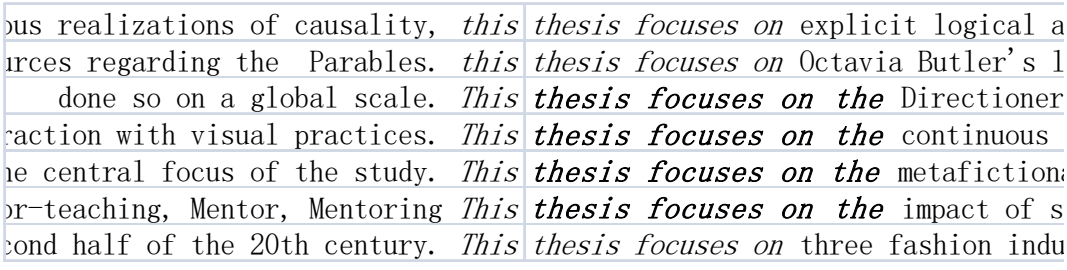

Figure 1. Example of overlapping lexical bundles

Table 3. 4-gram Lexical Bundles Retained in Two Sub-Corpora after the Removal of Overlaps

Corpora	*Types*	*Normalized per 10,000 words*	*Tokens*	*Normalized per 10,000 words*
CHTAs	103	10.5	1230	125.1
USTAs	37	5.7	239	36.8

2.3 Functional Classification

The functions of lexical bundles have received considerable attention in recent years, but there are some differences in the functional classification schemes used. Biber et al. (2004) used an inductive technique to group together expressions with the same function. They determined three primary functions: stances expressions, discourse organizers and referential expressions. However, they did note that "In some cases, a single bundle has

multiple functions even in a single occurrence" (Biber et al. 2004: 381).

Building on the work done by Biber et al. (2004), Hyland (2008b) developed functional classifications to fit the corpora used in his study. Biber et al used a corpus of spoken and written language, including conversations, service encounters and a variety of written texts. On the other hand, Hyland used written academic language, including research articles, doctoral dissertations, and master's theses. Hyland's taxonomy has three broad categories: research-oriented, text-oriented and participants-oriented. It should be noted that both Biber et al. (2004) and Hyland (2008b)'s functional categories contain a number of sub-categories.

Given that the corpora used in the current research most closely resembles those of Hyland (2008b), it was decided to use his functional taxonomy. That is, Biber et al. (2004) and Simpson and Ellis (2010) used both written and spoken language, including casual conversation, while Hyland (2008b) used only written academic language. However, when identifying the functions of the lexical bundles in the current study, we at times referred to the examples given by Durrant (2017). In his analysis of bachelor and master's level student writing, Durrant (2017) used Hyland's categories but made minor adjustments to accommodate the bundles in his study. Nevertheless, the extensive lists of bundles and their functions given by Durrant proved to be extremely helpful.

When actually using the functional categories, we were mindful of the caution from Adel and Erman (2012: 89). They have reservations about the use of functional classification, and claim that "No clear criteria are given for how to decide which (sub)category a given bundle should belong to". In the present study we made an effort to adhere to the functional criteria we did find, including the use of examples when possible. One researcher labeled the lexical bundles based on Hyland's criteria and referred to the examples in the studies of Biber et al. (2004), Simpson and Ellis (2010) and Durrant (2017) when there were questions. Then two researchers checked the classification together using the same methodology. At times, the sentence context of a lexical bundle would be considered. For multifunctional bundles, the predominant function was identified by checking their sentence contexts. Table 4 gives examples of the lexical bundles in each functional category. In section 3.3, we present the results of the functional analysis.

Table 4. Examples of Bundles Across the Functional Sub-Categories in CHTAs and USTAs

Category	*Sub-category*	*Examples*
Research-oriented	procedure	*the use of the, the analysis of, the way in which*
	location	*at the same time, at the end of, the beginning of the*
	quantification	*is one of the, one of the most, a large number of*
	description	*the characteristics of the, the structure of the, the main body of*
	topic	*the image of the*
Text-oriented	Structuring	*this thesis explores the, paper is divided into, the goal of this*
	framing	*as a form of, on the basis of, from the point of*
	resultative	*as a result of, the causes of the, the reasons for the*
	transition	*on the one hand, as well as how*
Participant-oriented	stance	*it is important to, are more likely to, is a kind of*

2.4 Structural Classification

Lexical bundles can be grouped into categories according to their grammatical types. Biber et al. (1999) distinguished 12 major structural categories in a 3.5-million-word corpus of academic prose. These 12 categories were reduced to three broad categories by Chen and Baker (2010) in their study using three corpora of expert academic prose, native student academic writing, and English language learner academic writing, each with about 150.000 words. Based on Biber et el's classification, lexical bundles in the Chen and Baker study were classified as "NP-based," PP-based," or one of six types of "VP-based" phrases. This classification system was used again by Biber in a 2016 study (Pan et al., 2016). For the present study, using only two smaller corpora, it was more appropriate to use these three broad categories with subcategories for the VP-based bundles.

The three broad structural categories were operationalized according to criteria used in Chen and Baker (2010). NP-based bundles included any noun phrases with post-modifier fragments, such as "the role of the" or "the way in which." PP-based bundles referred to those starting with a preposition plus a noun-phrase fragment, such as *at the end of* or *in relation to the*. With regard to VP-based bundles, any word combinations with a verb component, such as *in order to make* or *was one of the*, were assigned to this category. The six sub-categories of VP-based bundles included:

• Verb Phrase with active verb (this thesis focuses on; pay attention to)

• To-clause fragment (in order to explore)

• Noun phrase + be (the first chapter is)

• Copula be + noun/adjective phrase (is one of the; are more likely to)

• Anticipatory it + verb phrase/adjective phrase (it is important to)

• Passive verb +prepositional phrase fragment (can be divided into; is based on the)

In addition, there were a few 4-grams that did not seem to fit this classification scheme, *such as as well as the; as well as how;* and as *well as their*. These were coded as Others.

3. Results and Analysis

3.1 Frequency of Lexical Bundles in Two Sub-Corpora

The types and tokes of lexical bundles found in the two sub-corpora are given in Table 5. The Chinese corpus contained a greater number of both types and token than did the English corpus. While many studies on lexical bundles in corpus linguistics use basic descriptive statistics such as frequencies and percentages, in order to better understand these differences in frequencies between the two corpora, the log-likelihood test was used. The results indicate that the differences between the two corpora both for types and tokens are significant. This finding, that the non-native Chinese students used more lexical bundles than American students as well as a greater range of lexical bundles, differs from some previous studies. This finding will be discussed below.

Table 5. Types and Tokens of 4-gram Lexical Bundles Used by Chinese Students and American Students

	CHTAs (98,297 words)	*USTAs (65,021words)*	*Log-likelihood*	*P*	
Types	103	37	11.04	0.001***	+
Tokens	1230	239	384.38	0.000***	+

Note: The asterisks (*) indicate significance level, ($^{***}p<0.001$)

An examination of the most frequent lexical bundles in the two corpora also revealed differences. The ten most frequent lexical bundles used by each two group are in Table 6. Only one of these bundles, *as well as the,* was shared by the two groups of writers. Further examination of the data found that there were only four shared lexical bundles, *the purpose of this, as well as the, at the same time, the beginning of the.* The number of shared bundles only accounts for 2% of the total bundles of the two groups. This proportion is very small. Furthermore, the use of these shared lexical bundles was different. As seen in Table7, the Chinese students used a significantly greater number of one lexical bundle, *at the same time.* The native English students used a significantly greater number of another bundle, *the purpose of this.* Further discussion of these shared bundles is given below, but the lack of shared bundles points to another difference between the two groups.

Table 6. Ten Most Frequent 4-gram Lexical Bundles in CHTAs and USTAs

	CHTAs		*USTAs*	
	Bundles	*Raw Freq.*	*Bundles*	*Raw Freq.*
1	from the perspective of	79	the ways in which	24
2	at the same time	47	the purpose of this	21
3	In the process of	33	as well as the	15
4	as well as the	32	in this thesis i	11
5	is one of the	32	i argue that the	8
6	on the basis of	32	the goal of this	8
7	the analysis of the	32	this thesis explores the	8
8	the development of the	32	this thesis focuses on	7
9	purpose and significance of	31	as a result of	6
10	the fifth chapter is	24	as a way to	6

Table 7. Shared 4-gram Lexical Bundles in CHTAs and USTAs

Lexical bundles	*CHTAs (98,297 words) Frequency*	*USTAs (65,021word) Frequency*	*Log-likelihood*	*p*	
the purpose of this	7	21	14.30	0.000***	-
as well as the	32	15	1.26	0.262	+
at the same time	47	5	24.01	0.000***	+
the beginning of the	6	4	0.00	0.990	-

Note: The asterisks (*) indicate significance level, (***p<0.001) and the "+" and "-" signs on the right side indicate "overuse" and "underuse".

3.2 Structural Comparison of 4-Word Bundles in Two Sub-Corpora

After identifying the lexical bundles and comparing the frequency in both corpora, the lexical bundles were coded for their structural characteristics, using the three broad categories described above, noun-based (NP), preposition-based (PP), and verb-bases (VP) with the VP category divided into six sub-categories. Table 8 displays the distribution of lexical bundles across the structural categories in CHTAs and USTAs. For the CHTAs the distribution of both types and tokens is roughly equal across the three categories, NP-based bundles, PP-based bundles and VP-based bundles. Each category accounts for approximately one-third of the total.

A slightly different pattern is observed in USTAs. For the types, the PP-based bundles make up the largest structural category, accounting for about 38% with the remaining types about evenly divided between NP-based and VP-based bundles. For tokens, the NP-based and PP-based tokens are about equal, while VP-based bundles account for less than one-quarter of the tokens.

Table 8. Proportional distribution of Lexical bundles (types and tokens) across the structural categories in CHTAs and USTAs

	CHTAs		*USTAs*		*CHTAs*		*USTAs*	
Category	*%*	*Types*	*%*	*Types*	*%*	*Tokens*	*%*	*Tokens*
NP-based	35	36	24	9	33.7	414	33.9	81
PP-based	30	31	38	14	34.1	420	31.8	76
VP-based	34	35	27	10	29.6	364	22.2	53
Others	1	1	11	4	2.6	32	12.1	29
Total	100	103	100	37	100	1230	100	239

Some differences between the two corpora in the percentage of VP-based sub-categories can be seen in Table 9. With types, Chinese students used no *to*-clause fragments and anticipatory *it* +verb phrase/adjective phrase bundle, while these two sub-types together accounted for 30% of the types used by the US students. The US

students, on the other hand, used no passive verb + prepositional phrase bundles while this sub-type accounted for 14% of the bundles used by the Chinese students. The Chinese students also overused noun phrase + be bundles compared to the US students, 23% to 10%.

In regard to tokens, the Chinese students also used no to-clause fragment and anticipatory it + verb phrase/adjective phrase bundles. In addition, the Chinese students underused verb phrase with active verb bundle but overused noun phrase + be, and the US students used no passive verb + prepositional phrase fragment bundles.

Because of the relatively small number of types in each sub-category, the differences were not tested for significance. However, as seen above, the total types of VP-based bundles were significantly different, and an examination of the percentages revealed differences in the use of the sub-categories, but no meaningful patterns were evident.

Table 9. Proportional distribution of Lexical bundles (types and tokens) across VP-based sub- categories in CHTAs and USTAs

		CHTAs		*USTAs*		*CHTAs*		*USTAs*	
Category	*Sub-category*	*%*	*Types*	*%*	*Types*	*%*	*Tokens*	*%*	*Tokens*
VP-based	verb phrase with active verb	54	19	50	5	40.1	146	58.5	31
	to-clause fragment	--	--	20	2	--	--	17.0	9
	noun phrase + be	23	8	10	1	31.9	116	7.5	4
	copula be + noun/adjective phrase	9	3	10	1	12	46	9.5	5
	anticipatory it + verb phrase/ adjective phrase	--	--	10	1	--	--	7.5	4
	passive verb +prepositional phrase fragment	14	5	--	--	15.4	56	--	--
	Total	100	35	100	10	100	364	100	53

Table 10 presents the log-likelihood test results for significant differences between the structural categories in the two corpora for the types and tokens of bundles. In the three main categories, the Chinese students used a greater number of types, but the differences were significant only for NP-based and VP-based structural categories in the two corpora. The Chinese students used significantly more NP-based and VP-based bundle tokens than the American students. The Chinese students also used a greater number of tokens in all three main categories, and the differences were significant for the three categories.

Table 10. Log-likelihood test of lexical bundles (types and tokens) across the structural categories in CHTAs and USTAs

	Types			*Tokens*		
Category	*CHTAs*	*USTAs*	*p*	*CHTAs*	*USTAs*	*p*
NP-based	36	9	0.004^{**}	414	81	0.000^{***}
PP-based	31	14	0.226	420	76	0.000^{***}
VP-based	35	10	0.012^{*}	364	53	0.000^{***}
Others	1	4	0.066	32	29	0.222

Note: The asterisks (*) indicate significance level, ($^{*}p<0.05$, $^{**}p<0.01$, $^{***}p<0.001$)

3.3 Functional Comparison of 4-Word Bundles in Two Sub-Corpora

Despite the fact that functional classification is still problematic (Adel & Erman, 2012), we finalized the functional classification in the present study based on classification schemes used in previous research, examples in other studies, concordance lines of the current study, and extensive discussion.

In Table 11, it can be seen that the distribution of the three main functional categories, research-oriented, text-oriented, and participant-oriented, is very similar in the two corpora for both types and tokens. The greatest proportion in both corpora is text-oriented, and the smallest proportion is participant-oriented bundles.

Table 11. Proportional distribution of Lexical bundles (types and tokens) across the functional categories in CHTAs and USTAs

	CHTAs		*USTAs*		*CHTAs*		*USTAs*	
Category	*%*	*Types*	*%*	*Types*	*%*	*Tokens*	*%*	*Tokens*
Research-oriented	42	43	38	14	41.7	513	35.6	85
Text-oriented	57	59	54	20	57.8	711	57.3	137
Participant-oriented	1	1	8	3	0.5	6	7.1	17
Total	100	103	100	37	100	1230	100	239

While the percentage distribution of the functional categories was similar between the two corpora, a Log-likelihood test of the number of both the types and tokens shows that there are significant differences between the two corpora. As seen in Table 12 the Chinese students used significantly more research-oriented and text-oriented types and tokens. However, the Chinese students used significantly fewer participant-oriented tokens. There was no significant difference in participant-oriented types, but only a few types were used by both groups.

Table 12. Log-likelihood test of lexical bundles (types and tokens) across the functional categories in CHTAs and USTAs

	Types			*Tokens*		
Category	*CHTAs*	*USTAs*	*p*	*CHTAs*	*USTAs*	*p*
Research-oriented	43	14	0.015*	513	85	0.000 ***
Text-oriented	59	20	0.007 **	711	137	0.000 ***
Participant-oriented	1	3	0.153	6	17	0.001 ***

An examination of the functional subcategories reveals that the significant differences for types were confined to two sub-categories, description for the research category and structuring for the text category (Table 13). But the Chinese students tended to overuse all the sub-categories when compared to the US students, with the exception of transition bundles. In regard to tokens, the Chinese overused all of the subcategories when compared to the US students, and the differences were significant for all of the sub-categories with the exception of the transition sub-category.

Table 13. Log-likelihood test of lexical bundles across the functional sub-categories in CHTAs and USTAs

Functional Sub-categories	*CHTAs (98,297 words)* *No. of lexical bundles*	*USTAs (65,021word)* *No. of lexical bundles*	*Log-likelihood*	*p*	
Procedure	15	8	0.25	0.619	+
Location	9	4	0.46	0.498	+
Quantification	4	2	0.11	0.743	+
Description	13	0	13.20	0.003**	+
Topic	2	0	2.03	0.250	+
Structuring	33	8	7.77	0.005**	+
Framing	20	7	2.30	0.129	+
Resultative	3	1	0.39	0.533	+
Transition	3	4	0.85	0.356	-

The differences in the types found in the structuring sub-category were of particular interest. As seen in Table 14, a closer examination of the structural lexical bundles used in the USTAs revealed that they tend to refer to the whole, while many of those used by the CHTAs refer to a part. Furthermore, five of seven bundles used by the US students to refer to the whole are among the 10 most frequent bundles in the USTAs corpus, while none of the eight bundles used by the Chinese students to refer to the whole were among the 10 most frequent bundles in the CHTAs corpus. Only one of the 15 bundles used by the Chinese students to refer to the part is among the 10 most frequent bundles in the CHTAs corpus.

Table 14. Bundles in the sub-category of structuring in CHTAs and USTAs

Corpus	*CHTAs (Tokens)*	*USTAs (Tokens)*
Bundles	*** the fifth chapter is (24)**	**** the purpose of this (21)**
	* paper is divided into (22)	**** in this thesis I (11)**
	* the first chapter is (22)	**** the goal of this (8)**
	** this paper analyzes the (19)	**** this thesis explores the (8)**
	* the six chapter is (16)	**** this thesis focuses on (7)**
	* the second chapter is (15)	in order to explore (5)
	* the fourth chapter is (12)	** this thesis analyzes the (4)
	* can be divided into (11)	** this thesis is a (4)
	* the third chapter is (11)	
	* article is divided into (10)	
	* the first part is (10)	
	** this paper aims to (9)	
	introduces the research background (8)	
	is mainly about the (8)	
	**paper focuses on the (8)	
	the background of this (8)	
	**the paper analyzes the (8)	
	**this paper introduces the (8)	
	* chapter focuses on the (7)	
	the main content of (7)	
	**the purpose of this (7)	
	* thesis is divided into (7)	
	* third chapter introduces the (7)	
	the significance of the (6)	
	and purpose of the (6)	
	and the overall structure (6)	
	on the research of (6)	
	paper briefly introduces the (6)	
	the innovation of the (6)	
	* the second chapter describes (6)	
	* the second part is (6)	
	**this paper attempts to (6)	
	**this paper consists of (6)	

Note: Bold = item occurring in the list of ten most frequent bundles. * = item referring to the part. **=item referring to the whole.

One other difference found in the functional sub-categories is worth noting. There is no significant difference in the transition sub-category in regard to types. But on examination, it was seen that all four transition bundles in USTAs are based on one three-gram head "as well as" (Table 15). Only one transition sub-category used by the Chinese students contained this 3-gram, *as well as the.* Furthermore, the *as well as the* type was the most frequent transition bundle in both corpora. The resultative bundle *as a result of* used in the USTAs corpus was not represented in the CHTAs corpus. This last point will be discussed later.

Table 15. Bundles in the transition and resultative sub-categories in CHTAs and USTAs

Sub-categories	*CHTAs (tokens)*	*USTAs (tokens)*
Transition	as well as the (32)	as well as the (15)
	on the other hand (9)	as well as how (5)
	on the one hand (7)	as well as their (5)
		as well as to (4)
Resultative	the reasons for the (12)	as a result of (6)
	so that we can (6)	
	the cause of the (6)	

4. Discussion

In this section we will discuss the results of this study and compare them to other studies of L1 and L2 writers of English. However, it should be noted that comparison to other studies are tenuous at best due to variations in research design. These variations include different L1s of the writers, different genres used to construct the corpora, and different sections of academic writing in the same genre. Furthermore, the identification and classification of lexical bundles is problematic (Adel & Erman, 2012; Pan et al., 2016; Myles & Cordier, 2017). However, given this reservation, in this section after reviewing answers to the research questions, we will relate the findings to previous studies, and discuss the differences and similarities.

The analysis above provides answers to the research questions. It was found that

1) There were differences in the types and tokens of the lexical bundles used in the CHTAs and USTAs corpora. The Chinese students used significantly more types and tokens than did the US students.

2) There were differences in the structural characteristics of the lexical bundles found in the two corpora. The Chinese students used a significantly greater number of NP-based and VP-based types. The Chinese students also used a significantly greater number of tokens in all three categories.

3) In regard to the functional characteristics of both types and tokens, the percentage distributions were similar in both corpora, but there were significant differences in the actual count of the bundles. The Chinese students used a greater number of both types and tokens for research-oriented and text-oriented functions. However, the Chinese students used fewer participant-oriented tokens than did the US students.

4.1 Types and Tokens

Previous research has yielded conflicting results in regards to the number of types and tokens of lexical bundles used by non-native and native English speaking students. In their study of L1 Swedish university students, Adel and Erman (2012: 85) found that the native English writers produced a "considerably wider range of lexical bundles, that is types, than did the non-native writers. They went on to claim that the finding that "non-native speakers produce fewer and less varied lexical bundles" (Adel and Erman 2012: 85- that is, fewer tokens and types as "robust". Chen and Baker (2010) also found that non-native students produced fewer types and tokens of lexical bundles that did native students, and that the number of types and tokens increased with advancing proficiency. However, these observations are based on raw frequencies, not normalized data for the three corpora used in the study, all of which contained different numbers of words, with the L1 Chinese corpus being the smallest.

In spite of the assertion made by Adel and Erman (2012) that L1 writers produce a greater number of and more varied set of lexical bundles, there are a number of studies that have found just the opposite. The results of the present study are in accord with Güngör and Uysal (2016), Pan et al. (2016) and Hyland (2008a) on the finding that non-native speakers produced more lexical bundles than native speakers did in academic writing. As Paquot and Granger (2012:139) assert, "[L]earners tend to use more lexical bundles in writing when compared to native speakers .... Overall, less proficient learners seem to be more reliant on lexical bundles".

At first glance, it seems incongruous that L2 writers would use a greater variety and number of lexical bundles than L1 writers. A possible explanation is that because L1 writers have a great store of lexical bundles from which to choose, the use of a particular bundle fails to reach the cut-off level. On the other hand, L2 writers have a more restricted set of bundles from which to choose, and as a result make the same choices, leading to a greater number of bundles reaching the cut-off frequency. That is, as argued by Myles and Cordier (2017),

there is a wide variety of linguistic units which are processing units for l1 writers and thus readily available for use. The l2 writers had fewer processing units and had to resort to repeated use of those units, resulting in a greater number of linguistic units that reached cut-off frequency. Also, as Adel and Erman (2012) pointed out, a relatively high cut-off score, such as 60 per million used in the current research, "is likely to favor the learners who have a more restricted repertoire but tend to use their favorite bundles unusually often" (p. 88).

4.2 Structural Characteristics

The results of the structural analysis also corroborated the results of other studies. Academic writing is characterized by NP-based and PP-based bundles because of the informational load (Pan et al., 2016). In the present study, the percentage of NP-based and VP-based bundles accounted for nearly two-thirds of the types and tokens in both corpora. These results are also similar to those reported by, Hyland (2008a) and Conrad and Biber (2005) for academic writing. But the results in our study are contrary to the finding of Güngör and Uysal (2016) who reported that L1 Turkish writers used more VP-based bundles. These different results might be explained by the use of writing from different L1s, but Chen and Baker (2010) study with L1 Chinese students found that both native and non-native student essays contain many more VP-based bundles than native expert writing does. Clearly this is an area that needs more research with different L1 languages as well as different level of writers, including undergraduates, graduate students, and published expert writers.

Other factors may be at play as well. It was found by Shahriari (2017) that there were structural differences among the lexical bundles found in different sections of research articles. The research reported here used only abstracts, in contrast to the research of Güngör and Uysal (2016) who used entire research articles and that of Chen and Baker (2010), who used intact native and non-native student essays. Task demands may have also played a role in the results of the present study. Both corpora contain academic writing, abstracts for theses. However, for the Chinese abstracts, students were explicitly asked to summarize each part of the thesis. This could have resulted in the overuse of the VP-based sub-category, noun + copula, such as *the first part is.* Clearly this is an area that needs more research with different L1 languages as well as different language proficiency levels, greater specificity in the nature of the texts, and a consideration of task demands.

4.3 Functional Characteristics

Both similarities and differences were found in the functional uses of lexical bundles in the two corpora. Looking only at the percentage distributions, the functional use was similar for both types and tokes. When considering the actual count of the functional uses, differences were evident. Compared to the US students, the Chinese students used a greater number of both types and tokens for research-oriented and text-oriented functions, but the Chinese students produced few tokens of participant-oriented lexical bundles.

It is difficult to make direct comparisons of these findings with those of other studies because of the differences in functional classification schemes. Yet, in general it can be observed that other studies have had similar findings, as the proportional distributions were similar, but the actual count showed differences. For instance, Pan et al. (2016) also found that for both types and tokens, native and non-native expert writers did not differ much in the proportion distribution of functional bundles, but there were differences in the actual count of the tokens. The pattern of differences observed by Pan et al, however differed from our findings, as the L2 writers in the Pan et al study used significantly fewer research-oriented bundles and more stance-oriented bundles. The L2 students in our study used more research-oriented bundles and used fewer participant-oriented bundles, which are similar to stance-oriented bundles. Nevertheless, the similarities found in the current research concur with the general findings of Pan et al as well as those of Chen and Baker (2010), who concluded that "the use of lexical bundles in non-native and native student essays is surprisingly similar" (p. 44) for both types and tokens. The same general results were reported by Adel and Erman (2012) who found that the proportions of functional bundles between Swedish L1 writers and native English writers to be nearly the same, though they noted that the native speakers used a greater proportion of stance-bundles, but only by a few percentage points.

Contrary to these similar findings, other research did report conflicting results. When comparing Turkish L1 professionals writing in English with native professional writing, Güngör and Uysal (2016) found that the proportional distribution of research-oriented and text-oriented bundles was a near mirror image between the two groups. The percentage of research-oriented and text-oriented bundles for the Turkish L1 group was 69% and 28% respectively, while for the native English group the percentages were 30% and 66%. This is a contrast to our finding that text oriented bundles predominated for both the L1 and L2 writers. However, the two studies are based on different types of data. The data in the present study were abstract writing by native and non-native students while the data used in the study of Güngör and Uysal (2016) were research articles written by professional native and non-native scholars.

Other research has also found that writing expertise may be a factor in the functional uses of lexical bundles. Hyland (2008a) observed that for Cantonese L1 speakers writing in English, some differences appeared between PHD dissertations and master's theses. In the dissertations, over half of the types were text-based bundles and only about one-third were research-oriented bundles, while in the master's theses, the percentages were more equal. First language information was not given for the research articles, but 60% of the bundles were text-based, with research-based bundles account for only 25% of the bundles used in the research articles. In contrast, level of writing expertise was equal in the Pan et al. (2016) study. In their study, the corpora were composted of writing produced by L1 Chinese and native-English writers who were academic professionals published in telecommunications journals. The use of text-based bundles predominated, accounting for nearly 50% of the types in both corpora. However, in the current study the level of writing expertise for the undergraduate abstracts and master's abstracts was unequal, yet, the distribution of the functional types was nearly identical with text-oriented bundles predominating. Clearly the role of lexical proficiency in the use of lexical bundles needs further investigation.

While it is difficult to draw definitive conclusions comparing the present study to previous research because of differences in research design such as differences in functional classification schemes, L1, and writing expertise, one additional complicating variable may be educational context. The abstracts in the CHTAs corpus were written by undergraduate at a university in China that had particular requirements. They were longer, as the minimal length was 500 words. Furthermore, students were explicitly instructed to use text-oriented bundles such as *the first chapter.* Furthermore, they were translated into English from the original Chinese. Also, Chinese university instructors helped the students revise the abstracts. The extent of the aid given by the instructors is not known, but the issue is not unique to this study. For example, it is not known to what extent the academic professional writing in the telecommunications journals was revised by editors.

5. Conclusion and Implications

The impetus for the research reported in this article was to help the undergraduate students at a Chinese university use more natural language when translating thesis abstracts into English. While mindful of the native speaker fallacy (Hodgson, 2014), some recommendations can be made in regard to lexical bundles.

The most frequent lexical bundles found in the CHTA and the USTAs were quite different. The American students used a number of text-structuring bundles that helped signal the genre of abstract and gave the general area, the purpose of the thesis, and the results. Of the most frequent bundles used by the American students, five fulfill this function. The general area of the thesis is given using *in this thesis I, this thesis explores the,* and *this thesis focuses on.* The purpose was given by the use of two bundles, *the purpose of this,* and *the goal of this,* while *as a result of* was used to present the results.

Of the most frequent bundles used by the Chinese students, only one text-structuring bundle referred to the whole of the thesis by giving the purpose of the thesis, *purpose and significance of.* Another frequent text-structuring bundle used by the Chinese students, *the fifth chapter is*, refers to a part of the thesis, rather than the whole as did the frequent text-structuring bundles used by the American students. This bundle is typical of many text-structuring bundles used by the Chinese students, *the fifth chapter is, the first chapter is, the six chapter is, the first part is,* and *the second part is,* among others. These bundles were not used at all by the American students. One implication then, is that Chinese students be taught to avoid these text-structuring bundles that refer to the part and instead use more general text-structuring bundles referring to the whole. This is contrary to what is currently taught at the university, and may be a feature of abstract writing taught at other Chinese universities.

Another feature of the abstracts in the USTA corpus was the use of the 3-gram *as well as.* This 3-gram was frequently used as a part of the 4-grams, *as well as the, as well as how, as well as their,* and *as well as to* as a transition signaling bundle. The Chinese students used only one *as well as* transition signaling bundle, *as well as the.* In addition to this bundle, the Chinese students could be taught the other three transition *as well as* bundles. Simpson and Ellis (2010) and Martinez and Schmitt (2012) created a list of the most useful formulaic sequences for academic English, Academic Formulas List (AFL) and Phrasal Expressions List (PHRASE List) respectively. *As well as* is in both the lists. Moreover, Omidian, Shahriari, and Ghonsooly (2017) investigated the value judgments of experienced EFL teachers and advanced-level learners from AFL and PHRASE List and found *as well as* appeared on the most valuable top 10 list of both the teachers and the learners. Adding these functional uses should be relatively easy as the 3-gram, *as well as* is already known to the Chinese students. In general, we believe that instructors can help L2 students produce more coherent and more professional academic abstracts by improving awareness of lexical bundles and focusing on the use of the

specific lexical bundles in academic abstracts.

In addition to these pedagogical implications, this study provides some implications for the study of lexical bundles. First of all, it adds to the considerable body of research on the use of lexical bundles by second language writers. It has been noted that it is "particularly difficult to compare" this research because of different sized lexical bundles and varied settings for the data (Paquot & Granger 2012: 138). Comparison among studies is complicated by a number of important variables including, L1, L2 proficiency, general writing proficiency, type of writing, and context of writing. In particular, the researcher's experiences coding the data suggest that special care needs to be taken with the classification of functions. Another, less frequently considered variable is that of task demands, in the case of the present research, the requirements given to the Chinese students when writing the abstracts. It is expected that future research will address these variables.

References

Adel, A. & 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*, 81–92. https://dx.doi.org/10.1016/j.esp.2011.08.00

Alasmary, A. (2019). Academic lexical bundles in graduate-level math texts: a corpus-based expert-approved list. *Language Teaching Research, 00*, 1-25. https://dx.doi.org/10.1177/1362168819877306

Biber, D. & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., Conrad, S. & Cortes, V. (2004). If you look at...: lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., Johansson, S., Leech, G. & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Harlow: Pearson.

Chen, Y. H. (2009). *Lexical bundles across learner writing development* (Unpublished doctoral dissertation). Lancaster university.

Chen, Y. H. & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Teaching, 14*(2), 30–49. https://doi.org/ 10.1111/j.1467-9922.2009.00559.x

Conrad, S. M. & Biber, D. (2005). The frequency and use of lexical bundles in conversation and academic prose. *Applied Linguistics, 10*(1), 55–71. https://doi.org/ 10.1515/9783484604674.56

Durrant, P. (2017). Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. *Applied Linguistics, 38*(2), 165–193. https://doi.org/10.1093/applin/amv011

Grabowski, L. & Juknevičienė, R. (2016). Towards a Refined Inventory of Lexical Bundles: an Experiment in the Formulex Method. *Studies About Languages, 29*, 58–73.

Güngör, F. & Uysal, H. H. (2016). A comparative analysis of lexical bundles used by native and non-native scholars. *English Language Teaching, 9*(6), 176–188. https://doi.org/ 10.5539/elt.v9n6p176

Hodgson, K. M. (2014). Mismatch: Globalization and Native Speaker Models of Linguistic Competence. *RELC Journal, 45*, 113–134. https://10.1177/0033688214533863

Hyland, K. (2008a). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62. https://dx.doi.org/10.1111/j.1473-4192.2008.00178.x

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purpose, 27*, 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K. (2012). Bundles in Academic Discourse. *Annual Review of Applied Linguistics, 32*, 150–169. https://doi.org/10.1017/S0267190512000037

Martinez, R. & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299–320. https://doi.org/ 10.1093/applin/ams010

Myles, F. & Cordier, C. (2017). Formulaic sequence(fs) cannot be an umbrella term in SLA: Focusing on psycholinguistic FSs and their identification. *Studies in Second Language Acquisition, 39*, 3–28. https://dx.doi.org/10.1017/S027226311600036X

Nation, I. S. P. & Webb, S. (2011). *Researching & Analyzing Vocabulary*. Heinle: Cengage Learning.

Omidian, T., Shahriari, H. & Ghonsooly, B. (2017). Evaluating the Pedagogic Value of Multi-Word Expressions

Based on EFL Teachers' and Advanced Learners' Value Judgments. *TESOL, 8*(2), 489–511. https://doi.org/ 10.1002/tesj.284

Pan, F., Reepen, R. & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in telecommunications research journals. *Journal of English for Academic Purposes, 21*, 60–71. https://doi.org/10.1016/j.jeap.2015.11.003

Paquot, M. & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130–149. https://doi.org/ 10.1017/S0267190512000098

Shahriari, H. (2017). Comparing lexical bundles across the introduction, method and results sections of the research article. *Corpora, 12*(1), 1–22. https://doi.org/ 10.3366/cor.2017.0107

Shin Y. K. (2018). The construction of English lexical bundles in context by native and nonnative freshmen university students. *English Teaching, 73*(3), 115-139. https://doi.org/10.15858/engtea.73.3.201809.115

Simpson, R. & Ellis, N. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 485–512. https://doi.org/10.193/applin/amp058