THE ASIAN ESP JOURNAL

Sentence Initial Bundles: A Comparative Study Between Chinese Master's L2 Theses and Published Writing

Liang Li, Margaret Franken, Shaoqun Wu

University of Waikato, New Zealand

Bio Data

Liang Li is a research assistant in the Faculty of Education. She has just completed her PhD, comparing sentence initial bundles in Chinese L2 and New Zealand L1 thesis writing. Her research interests lie in the area of corpus linguistics, L2 academic writing, and computer-assisted language learning.

Email: liang.li@waikato.ac.nz

Margaret Franken is an associate professor in the Faculty of Education. She has had a long-term research interest in supporting second language writers at tertiary level. She is the co-author of a number of articles evaluating the way in which corpora and digital library type's resources can do this.

Email: margaret.franken@waikato.ac.nz

Shaoqun Wu is a senior lecturer at Faculty of Computing and Mathematical Sciences. Her research interests include computer assisted language learning, mobile language learning, supporting language learning in MOOCs, digital libraries and natural language processing.

Email: shaoqun.wu@waikato.ac.nz

Abstract

Lexical bundles, like recurrent multi-word combinations, act as discourse frames in a register and so are potentially significant as markers of expertise. The present study compared sentence initial lexical bundles (i.e. bundles at the beginning of sentences) in 43 Chinese Master's theses written in English and 85 published research articles written by L1 or advanced L2 writers of English in terms of their frequency, grammatical structures and related discourse functions. The Chinese Master's L2 texts showed a number of distinctive features, including but not restricted to an overuse of general nouns, pronoun *it* and sentence connectors, and an absence of shell nouns, anticipatory-*it* and some less transparent bundles. This paper discusses some of the possible reasons for these findings and indicates a need for pedagogic attention to cohesive devices and salient bundles which can be implemented with the help of effective corpus-based tools.

Keywords: Lexical bundles; Chinese students; Academic writing; Corpus analysis

Acknowledgements

We would like to express our sincere gratitude to the editors and reviewers for their careful and constructive comments and suggestions.

1. Introduction

Lexical bundles, as recurrent multi-word combinations, are identified on the criterion of distribution as they have a high frequency of occurrence and wide distribution across texts (Biber, Johansson, Leech, Conrad, & Finegan, 1999). These combinations are extremely common discourse building blocks in a given register, and they act as discourse frames to connect to new information (Biber & Barbieri, 2007) or as interactional devices for the involvement of the writer and engagement of target readers (Hyland, 2005, 2008c). As an effective approach to corpus-based analysis, lexical bundles have attracted an increasing number of studies in the last decade. Bundles have been investigated in relation to their use in different languages (e.g. Kaneyasu, 2012; Kim, 2009; Tracy-Ventura, Cortes, & Biber, 2007), different registers (e.g. Biber, 2006; Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Herbel-Eisenmann & Wagner, 2010; Jablonkai, 2010; Neely & Cortes, 2009; Nesi & Basturkmen, 2006; Schnur, 2014) and different genres (particularly of English) (e.g. L. Chen,

2010; Hyland, 2008a; Qin, 2014; Xu, 2012). Their role in accounting for general language proficiency as well as specific academic competence has also been explored (e.g. Ädel & Erman, 2012; Allen, 2009; Li, 2016; Y.-H. Chen & Baker, 2010; Hyland, 2008a; Staples, Egbert, Biber, & McClair, 2013; Wei & Lei, 2011).

The methodology of studies in these areas is generally similar. Four words are regarded as the most appropriate length for target bundles because these clusters present a wider range of structures than three-word clusters and recur more regularly than five-word clusters (Hyland, 2008b). The frequency threshold is normally 10 to 25 times per million words across 3 to 5 texts (e.g. Ädel & Erman, 2012; Biber et al., 1999; Y.-H. Chen & Baker, 2010; Cortes, 2004). However, some researchers take a relatively conservative approach to manipulating their data to a manageable size, setting the cut-off frequency as 40 times per million words (e.g. Pan, Reppen, & Biber, 2016) across 10% of texts (e.g. Hyland, 2008a, 2008b), or 10 to 20 texts (e.g. Wei & Lei, 2011).

Structural analysis has been an important focus of nearly all studies. In academic prose, noun phrases (e.g. *the use of the*) and prepositional phrases (e.g. *in the present study*) comprise over 60% of all lexical bundles (Biber, Conrad, & Cortes, 2003; Biber et al., 2004; Biber et al., 1999). Together with passive verb phrases (e.g. *can be found in*) and anticipatory-*it* patterns (e.g. *it is important to*, *it was found that*), these four structures are the most common patterns of lexical bundles in academic writing (Hyland, 2008a). The following discussion mainly reports on the identified differences between L2 writing (particularly Chinese learner writing), and native or expert writing with regard to these four patterns.

Noun phrase bundles, mostly with an embedded *of*, were found to occur more frequently in essays written by native speakers or in journal articles. Chinese undergraduates and Master's students do not appear to recognise the importance of this structure (Chen & Baker, 2010; Hyland, 2008a; Pang, 2009; Xu, 2012). However, in comparison to Chinese undergraduate and Master's writing, the distribution of noun phrase bundles in Chinese PhD writing tends to be closer to their distribution in published writing (Qin, 2014; Wei & Lei, 2011; Xu, 2012).

The use of prepositional phrase bundles in Chinese student writing has been found to increase with their levels of study. At the undergraduate level, Chinese students have been found to use considerably fewer bundles than native writers (Pang, 2009). From the undergraduate to Master's

level, they have been shown to employ a similar proportion of PP-based bundles to native and expert writers, slightly higher than their native peers but lower than expert writers (Chen & Baker, 2010). At PhD level, they appear to rely more heavily on PP-based bundles in comparison to Master's students and expert writers (Hyland, 2008a). It is possible that the students with higher degrees are more likely to be expected to construct longer texts, which may require a wider range of PP-based bundles to elaborate logical connections between units of texts (e.g. *on the other hand*) or to specify pre-conditions of their arguments (e.g. *on the basis of*).

Passive verb bundles were rarely found in Chinese and Swedish L2 university writing (Ädel & Erman, 2012; Chen & Baker, 2010), but were frequent in Chinese students' Master's and PhD theses (Hyland, 2008a; Wei & Lei, 2011). The use of anticipatory-*it* structures also differs across studies. Hyland (2008a), and Ädel and Erman (2012) found that anticipatory-*it* patterns were more common in Hong Kong and Swedish students' writing than in that of journal article writers. In contrast, Xu (2012), and Wei and Lei (2011) report that Chinese learners employ fewer anticipatory-*it* structures than writers of published articles. Differences between the Chinese students' writing and native or published writing are also evident in the use of *to*-clause fragments. Chinese undergraduates show a strong preference for *to*-clause fragments, especially the structure (*in order) to + verb* (Chen & Baker, 2010; Pang, 2009).

The above studies provide a justification for investigating and teaching lexical bundles. In these studies, published articles have often been used as a model of writing to explore the divergence of student bundle production. However, many studies did not consider overlaps between bundles. While generating bundles, a corpus tool reads from the first word of each text in the corpus and advances one word at a time. Along with the reading process, the tool stores every n-word sequence (i.e. n-gram) and checks against its previously identified sequences. Therefore, the generating process is highly likely to result in bundle overlaps. Y.-H. Chen and Baker (2010) suggest that bundle overlaps consist of complete overlap and complete subsumption. Complete overlap happens when two short bundles are both part of a long one. As they illustrated, both 4-word bundles *it has been suggested* and *has been suggested that* came from the 5-word one *it has been suggested that* and occurred with the same frequency in their corpus. Complete subsumption refers to the situation that one short bundle occurs more times than another, but both are subsets of a longer one. For example, the bundle *as a result of* was more frequent than *a result of the* in their

corpus, but both bundles were subsets of the 5-word bundles *as a result of the*. Both types of bundle overlaps will inflate the results of quantitative analysis and lead to an inaccurate comparison.

Drawing on the previous research, we aim to focus on sentence initial bundles (i.e. bundles at the beginning of sentences) in the present study. The focus on sentence initial bundles not only avoids the time-consuming and painstaking process of manually checking bundle overlaps, particularly in a larger corpus, but also reveals the specific function of sentence starters. As Cortes (2013) argues in her research, sentence initial and non-initial bundles function differently as *triggers* and *complements*: the former overlap with themes of sentences (Flowerdew, 2013; M.A.K. Halliday & Matthiessen, 2004) and function as the departure point of messages to locate and orient the clauses (e.g. *It should be noted*), while the latter act as complements to complete clauses or provide additional information (e.g. *the extent to which*). According to Williams (2003) and Hinkel (2004), starting a sentence is more challenging for writers, as this demands both sequencing the subsequent information and meeting the reader's expectations.

In this study, we compare the use of sentence initial bundles in Chinese Master's L2 theses and published research articles. Published research articles were used as a model of writing to identify the divergence of learner bundle production. We randomly selected the texts in the domain of corpus-based lexical analysis to identify the differences and similarities between the bundles written by Chinese students and published writers. The following research questions were developed to focus our investigation:

1. What are the most frequent sentence initial 4-word combinations (i.e. sentence initial bundles) in the two corpora?
2. How are these sentence initial bundles classified structurally?
3. To what extent do the sentence initial bundles employed by student writers differ from those of published writers?
4. What might be the potential reasons for the differences?

On the basis of our findings, some pedagogical implications will also be explored.

2. Corpora and Methods

The present study is based on a learner corpus of Chinese Master's L2 theses and a published research article corpus, both built within the domain of corpus-based lexical analysis. Although the two genres differ in length, audience and purpose, they both "represent the key research genres of the academy" (Hyland, 2008a, p. 47) with shared moves of research genres. Following the practice of many bundle studies (e.g. Y.-H. Chen & Baker, 2010; Cortes, 2004; Hyland, 2008a; Wei & Lei, 2011), the research articles, "as a model of good academic writing and as an ideal to be emulated" (Hyland, 2008a, p. 47), are used in this study to reveal the divergence of learner bundle production between Chinese Master's students and academics, that is, "the potential disparity between established characteristics of published writing and L2 writing" (Crawford, 2008, p. 269). The Chinese Master's L2 these corpus and published research article corpus were built within the same domain because Tse and Hyland (2009) suggest that the analysis of only one type of text in just one specific domain can be more effective for pedagogy than analyses of general academic English. This study consists of three stages, namely, corpus collection, bundle identification and bundle analysis.

2.1. Corpus collection

To build the learner corpus, we downloaded 43 Chinese Master's theses, written in English, totalling 839,922 words, from Wanfang Data Knowledge Service Platform. These theses were written by Chinese Master's students of 31 universities in mainland China and published between 2000 and 2012. The students are English majors who have been learning English for at least thirteen years. Their English proficiency can be considered as upper-intermediate to advanced level (i.e. above IELTS 5). It should also be noted here the collected theses are mostly likely to have been revised by the supervisors since published theses are final products of a Master's degree.

To compose the published corpus, we randomly collected 85 relevant research articles, published from 2000 to 2012, totalling 521,259 running words, from 42 different English-medium peer-reviewed journals using the leading research databases — Cambridge Journals Online, EBSCOhost Megafile Premier, and ScienceDirect (Elsevier). The authors of these articles are from 19 countries: about half from English-speaking countries (e.g. the UK, the USA and New Zealand) and another half from non-English-speaking countries (e.g. Iran, Italy, P. R. China, Sweden and

Japan). It is assumed that these articles are representative of high standards because they were all collected from peer-reviewed journals.

2.2. Bundle identification

FLAX (http://flax.nzdl.org), a self-access language learning and analysis system, documented in Wu, Franken and Witten (2009; 2010), was used in this study. FLAX can automatically generate four-word lexical bundles from corpora, and categorise the retrieved bundles into sentence initial and non-initial ones in terms of their positions — at the beginning or in the middle of the sentences, making it a useful tool for this study.

In FLAX, the frequency and distribution threshold is pre-set as 3 occurrences across 3 texts to avoid individual author idiosyncrasies. In the literature, the frequency threshold usually ranges between 10-40 times per million words and the distribution threshold is at least 3-5 texts (e.g. Ädel & Erman, 2012; Y.-H. Chen & Baker, 2010; Cortes, 2002, 2004, 2013; Hyland, 2008a, 2008b; Wei & Lei, 2011). In this study, as a result of the distinction between sentence initial and non-initial bundles, we used a less conservative threshold against the size of the corpora and the occurrence of the sentence initial bundles: the cut-off frequency is 5 times for the learner corpus and 3 times for the expert corpus, that is, 6 times per million words for both corpora. The distribution is at least 3 texts. This frequency is comparatively lower than the cut-off points in the literature (i.e. 10+ times per million words). However, a lower cut-off point is usually established for less common bundles. For example, Biber et al. (1999) set 5 times per million words for 5-word and 6-word bundles, and Cortes (2013) chose 8 times per million words for 6-word and 7-word bundles and 6 times per million words for her longer ones. Like longer bundles, sentence initial bundles are less common bundles, so they also deserve a less conservative cut-off point, that is, 6 times per million words in this study.

Content-based bundles (e.g. *The following concordance lines*), the bundles in the headers (e.g. *Available online at www.sciencedirect.com*), footers (e.g. *Further reproduction prohibited without*), acknowledgements (e.g. *We would also like*) and references (e.g. *Paper presented at the*), were manually removed from the data. As a result of the removal, 35 student bundles and 46 published bundles were discarded. Due to the domain-specific content of the texts, more content-based bundles (e.g. *The following concordance lines*) were discarded in this study compared with

the previous research on general or discipline-specific corpora (e.g. Ädel & Erman, 2012; Chen & Baker, 2010).

2.3. Bundle analysis

In the present study, the structural types and patterns were developed from Biber et al. (1999, 2004), and Chen and Baker (2010). On the basis of the Longman Spoken and Written English Corpus, Biber and his colleagues identified twelve widely-used structural patterns in academic prose, which are:

1. noun phrase with *of*-phrase fragment
2. noun phrase with other post-modifier fragment
3. prepositional phrase with embedded *of*-phrase fragment
4. other prepositional phrase fragment
5. anticipatory it + verb phrase/adjective phrase
6. passive verb + prepositional phrase fragment
7. copula be + noun phrase/adjective phrase
8. (verb phrase +) *that*-clause fragment
9. (verb/adjective +) *to*-clause fragment
10. adverbial clause fragment
11. pronoun/noun phrase + be (+ …)
12. other expressions

Biber et al. (2004) later developed three broad structural categories to group their structural patterns featuring in conversation, university teaching, textbooks and academic prose. These categories were bundles incorporating verb phrase fragments, dependent clause fragments and noun or prepositional phrase fragments. Along with Biber et al. (2004), but only focusing on academic writing, Chen and Baker (2010) distinguished another three major categories: noun phrase based (NP-based), preposition phrase based (PP-based) and verb phrase based (VP-based) bundles.

With reference to the categories of Biber et al. (2004); Biber et al. (1999), and Chen and Baker (2010), the first two authors of this paper worked independently to code a proportion of about 20% of the data and the inter-coder reliability was around 97%. Disputed cases on coding were resolved

during discussions. Then the first author coded the rest of the data, and codes were double-checked and refined by the other two authors.

Four major groups of bundles were identified: NP-based, PP-based, VP-based and clause-based bundles. In addition, two new patterns were created, *noun phrase + verb phrase* and *conjunction + clause fragments*, as a result of the sentence initial and non-initial bundle division. Table 1 gives the examples of each pattern. NP-based bundles refer to any noun phrases with post-modifier fragments, such as *of*-phrase fragments, post-nominal clause fragments, or any other preposition phrase fragments. PP-based bundles are preposition phrases or preposition phrases plus noun phrase fragments. VP-based bundles are composed of verb phrase fragments, *(In order) to-clause fragments* in this study. Clause-based bundles begin with independent or dependent clauses, and here refer to *anticipatory it-clause* fragments and the two newly-developed patterns: *noun phrase + verb phrase* and *conjunction + clause fragments*.

Table 1: Major Structural Categories and Patterns

Categories	Patterns		Examples
NP-based	noun phrase with post-modifier fragment	of	*The results of the*
		other	*The fact that the*
PP-based	preposition + noun phrase fragment	of	*On the basis of*
		other	*On the other hand*
VP-based	(verb/adjective) +	to-clause fragment	*In order to make*
Clause-based	anticipatory it +	VP	*It was found that*
		adjectiveP	*It is important to*
	noun phrase +	VP	*The results showed that*
	conjunction +	clause fragment	*As can be seen*
Other	other expressions		*That is to say*

3. Results and discussion

Regarding research question 1, we identified a total of 91 bundles in the learner corpus and 70 bundles in the published corpus. Appendix presents comprehensive lists of bundles identified in

the two corpora. Eight out of ten learner-preferred bundles rarely occurred in the published corpus. To address research question 2, Table 2 presents a comparison of the structural distribution of the bundles between the two corpora in terms of both type and token. In response to research question 3, log-likelihood tests were conducted using Paul Rayson's calculator (http://ucrel.lancs.ac.uk/llwizard.html). The results show that the journal article writers used significantly more NP-based bundles and *anticipatory it + adjective* bundles. The Chinese Master's students used significantly more VP-based bundles, *anticipatory it + verb* bundles, *noun + verb* bundles and other bundles. To better answer research question 3 and 4, the following sections will address the differences between bundles used in the Chinese Master's writing and published writing, and explore the possible reasons. Some typical bundles of Chinese student writing will also be discussed.

Table 2: Distribution of Sentence Initial Bundles by Structure (types and tokens)

Categories	Patterns		Types		Tokens	
			theses	articles	theses	articles
NP-based	noun phrase with post-modifier fragment	of	7	13*	46	46*
		other	2	3	11	16*
PP-based	preposition + noun phrase fragment	of	10	11	122	65
		other	9	10	132	89
VP-based	(verb/adjective) +	to-clause fragment	6	0*	46	0**
Clause-based	anticipatory it +	VP	12	2*	102	16**
	anticipatory it +	adjectiveP	2	8**	17	45**
	noun phrase +	VP	28	16	175	65**
	conjunction +	clause fragment	8	4	54	21
Other	other expressions		7	3	90	12**
Totals			91	70	795	375**

* = significant at $p < .05$ level ** = significant at $p < .01$ level

3.1. NP-based bundles

According to Cortes (2013), most nouns in these bundles are shell nouns. Shell nouns are also known by various names: general nouns (M. A. K. Halliday & Hasan, 1976), anaphoric nouns (Francis, 1986), carrier nouns (Ivanič, 1991), enumerative nouns (Hinkel, 2001, 2002, 2004) signalling nouns (Flowerdew, 2003), stance nouns (Jiang & Hyland, 2015) and meta-discursive

nouns (Jiang & Hyland, 2016, 2017). These nouns are pervasive in academic discourse, and carry little or no meaning, but operate to encapsulate the meaning from the preceding and succeeding clauses or noun phrases. Aktas and Cortes (2008) found shell nouns could serve a characterisation function (e.g. the *problem* of this technique), a temporary concept-formation (e.g. the same *result*), and a linking function (e.g. this *fact*) in academic prose.

As shown in Table 3, the journal article writers used a relatively wide range of shell nouns as the subjects in the pattern *the* + *N* + *of* to characterise and anticipate the *results/findings*, *analysis*, *aim/purpose*, *reasons*, and *design* of their studies or the *use* of various methods, whereas the student writers rarely deployed these shell nouns, except for *results*. The other two shared shell nouns, *size* and *number*, were used to describe corpora (e.g. *The size of the corpus*) or corpus data (e.g. *The total number of collocations*). This is because we built the two corpora within the same domain of corpus-based lexical analysis and the introduction of the size of a corpus and the number of generated data is crucial for corpus research.

Table 3: NP+*of* bundles

Student bundles	Frequency	Published bundles	Frequency
The results of the*	11	**The results of the**	10
The second type of	11	The results of this	10
One of the most	10	The analysis of the	8
The examples of the	6	The aim of this	8
The range of the	6	The findings of this	8
The size of the	6	The first of these	6
The total number of	6	The findings of the	6
		One of the reasons	6
		The purpose of this	6
		The size of the	6
		The total number of	6
		The design of the	6
		The use of these	6

*Sentence initial bundles in bold are shared bundles. Considering the two corpora were of different size, the final frequencies were normalized to 1,000,000 words to conduct a reliable comparison.

Another interesting finding is that nearly half of the *NP* + *of* bundles in the published writing ended with demonstrative determiners, *this* or *these*, as in:

The results of this *study are intended to be used with beginning and low intermediate learners of English whose vocabulary size is around 1,000 words. (published corpus, determiners)*

The extensive use of these two determiners, with an immediate referential function, are claimed to enhance the textural cohesion of academic writing (Biber et al., 1999; Halliday & Hasan, 1976; Hinkel, 2004). However, no demonstratives were found in the student *NP + of* bundles.

In the case of *NP + other modifications*, the nouns of the published bundles, *The **fact** that the*, *One possible **explanation** for*, and *The first **step** in*, were also used as shell nouns, as in the following examples.

The fact that the *learners in group 3 have spent more time in the target language community probably means that they have been exposed to more input generally. (published corpus, shell noun)*

One possible explanation for *these historical developments is the unique production circumstances of writing, which permit extensive planning and revision, in contrast to the real-time production circumstances of speech. (published corpus, shell noun)*

The first step in *the analysis was to identify all recurrent multi-word sequences in these two corpora. (published corpora, shell noun)*

In contrast, the nouns in the student bundles, *The **information** such as*, and *One **thing** to be*, are vague nouns, as in:

The information such as *the level of students, sex, age, school, the nature or source of the assignment, the category of genre of the writing and even the information about whether dictionary is used at the time of writing are also recorded in the entries of the data collected. (student corpus, vague noun)*

One thing to be *pointed out is that there is no clear-cut point for distinguishing free combinations, collocations and idioms. (student corpus, vague noun)*

Vague nouns are generic nouns, used to convey generalisation (Quirk, Greenbaum, Leech, & Svartvik, 1985; Sinclair, 1991). This finding is consistent with the findings of Hinkel's (2002,

2004) studies on the overuse of vague nouns in L2 university student writing. However, our finding of NP-based bundles highlights the role of shell nouns and demonstrative determiners in terms of sentence initial bundles, that is, as recurrent sentence starters.

3.2. PP-based bundles

A preliminary analysis of PP-based bundles showed that most bundles were complex prepositions (Hinkel, 2004), that is, multi-word preposition sequences used to clue texts (e.g. *based on*, *in spite of*, and *in addition to*). The student writers and their professional counterparts showed similarity in their choice of complex prepositions with many overlapping preposition bundles in the two corpora (see Table 4, the bundles in bold). Three out of the top five bundles in the pattern *PP + of* (*On the basis of*, *With the help of*, and *In the case of*) and all top three bundles of *PP + other modifications* (*In the present study*, *On the other hand*, and *At the same time*) were the same, although they were not sequenced in the exactly same order. The master level students appear to be comparatively competent to employ preposition units to join their ideas. One possible reason is these preposition units as common cohesive devices are often covered in writing courses, which are taught very early on and are frequently used during writing.

Table 4: PP-based Bundles

Student bundles	Frequency	Published bundles	Frequency
PP + *of* bundles			
On the basis of	29	**On the basis of**	29
With the help of	23	**In the case of**	25
In the process of	23	In terms of the	13
With the development of	20	**With the help of**	12
In the case of	12	As a result of	10
As a matter of	10	For the purposes of	8
With the popularization of	8	For the purpose of	6
In view of the	8	In the majority of	6
In spite of the	7	With the exception of	6
In one of his	6	In light of the	6
		In their study of	6
PP + other modification bundles			
In the present study	51	**On the other hand**	69
On the other hand	38	**At the same time**	23
At the same time	21	**In the present study**	21
On the one hand	12	**On the one hand**	15

In addition to the	8	**With regard to the**	10
In the same way	7	In this article we	8
In the following section	7	**In addition to the**	8
In the following part	6	Despite the fact that	6
With regard to the	6	In addition to these	6
		In the first part	6

However, it is important to note the prevalence of *with* bundles (*with the development of* and *with the popularization of*) in the Chinese students' writing. The use of these *with* bundles may originate in the interlingual transfer from the equivalent Chinese expression *随着,* and the Chinese students tend to be familiar with this pattern.

Two journal article-preferred idiomatic bundles, *In terms of the* and *In light of the*, were absent in the student writing. Both were used to provide the topic or theme of the arguments, as in:

> ***In terms of the*** *occurrence of referential bundles, it was found that they are more common in conversation than academic prose in both Korean and Spanish, which differs from English lexical bundles. (published corpus, idiomatic bundle)*
> ***In light of the*** *precision obtained from the last section, even though 94.1% of suggestions contain the appropriate corrections, no evidence shows whether our system can provide the most relevant answers with better ranking or not. (published corpus, idiomatic bundle)*

One possible explanation for the absence of these two bundles in the student corpus is the intransparency of the idiomatic expressions — the meanings cannot be interpreted from the literal meanings of their content words. In contrast, more transparent phrases such as *on the basis of*, *in the case of* and *with the help of*, were pervasive in the student texts. It is possible that the students found little or no difficulty in using the transparent expressions, but were not familiar with the less transparent ones and not confident in using them. However, according to Pawley and Syder (1983) the absence of a nativelike selection is likely to hinder the decoding process and increases the reader's processing load.

3.3. VP-based bundles

VP-based bundles were only found in the Chinese Master's corpus and the Chinese students habitually used *In order to* or *to*-clusters (*In order to make*, *In order to get*, *In order to have* and *To put it in*) at the beginning of their sentences to highlight the purposes of their main clauses. However, none of these patterns occurred as sentence initial bundles in the published writing, although the academics employed some in the second part of their sentences. The difference can be seen from the following examples:

> ***In order to make*** *up for the vocabulary deficiency, Chinese EFL learners tend to adopt repetition of some verbs they assume they are familiar with and avoid some verbs that they felt to be difficult as strategies of communication. (student corpus, in order to)*
> *This also shows how important it is for language learners to acquire a large number of phraseologies and patterns* ***in order to be*** *admitted into a discourse community, the wish to blend in. (published corpus, in order to)*

The use of sentence initial (*in order*) *to*-clusters may be attributed to the transfer of the Chinese phrase 为了, which usually starts a Chinese sentence. As Williams (2003) points out, long introductory phrases hinder understanding and readers "have to hold in mind that the subject and verb of the main clause are still to come" (p. 138). Therefore, it is more appropriate to start a sentence with its topic rather than the wordy (*in order*) *to* phrase in most cases.

3.4. Clause-based bundles

Clause-based bundles consist of anticipatory-*it* bundles, *noun + verb* bundles, and conjunction bundles. Among them, anticipatory-*it* bundles were heavily used in both student and published writing.

3.4.1. Anticipatory-it bundles

Anticipatory-*it* bundles were common in this research. The students employed more bundles in the pattern of *It + (modal) + passive verb + that* (i.e. 86%), but the journal article writers used more in the structure of *It + is + predictive adjective + to/that* (i.e. 87%). The use of anticipatory-*it* allows the writer to depersonalise the text and at the same time to take action against or make an

evaluation of the proposition; the use of anticipatory-*it* also accords with the information principle — new or heavy information is usually located at the end of the clauses (Williams, 2003). Hewings and Hewings (2002) identified four categories on the basis of the metadiscousal functions of their anticipatory-*it* data: emphatics (emphasising the writer's conclusion or drawing the reader's attention to a particular point), hedges (indicating the writer's uncertainty), attitude markers (expressing the writer's evaluation), and attribution (presenting the specific or general reference). Table 5 shows the distribution of the student and published anticipatory-*it* bundles in each category.

Table 5: Anticipatory-*it* Bundles

	Student bundles	**Frequency**	**Published bundles**	**Frequency**
Multi-functions	It can be seen	20		
Emphatics	**It was found that**	17	**It was found that**	15
	It is found that	17	It should be noted	15
	It is obvious that	13	It is important to	19
	It is hoped that	11	**It is clear that**	13
	It is also found	10	**It is obvious that**	6
	It is clear that	7	It is also important	6
	It can be inferred	7		
	It is expected that	7		
	It shows that the	6		
	It turns out that	6		
Hedges			It is possible that	12
				12
Attitude markers			It is difficult to	
			It is interesting to	12
			It is also worth	8
Attribution	It is suggested that	8		
	It is known that	7		
	It is believed that	6		

The most frequent bundle in the student corpus was *It can be seen*, which did not occur in the published corpus, but fulfilled a wide variety of functions as a discourse organiser, emphatics, and attribution in the student writing, for examples:

It can be seen that *in the above lines, there are some modifiers between "learn" and "lesson". (student corpus, discourse organiser)*

It can be seen that *the adjective-noun collocational errors is common among the four groups of students. (student corpus, emphatics)*

It can be seen that *Firth's definition and explanation of collocation lays a theoretical foundation for further research. (student corpus, attribution)*

All the shared bundles between the two corpora, *It was/is (also) found that*, *It is obvious that* and *It is clear that*, fell into emphatics category, used to state the writer's conclusions or deductions, for example:

It was found that *in both tests only around half of the students' responses were acceptable English collocations. (student corpus, emphatics)*

It was found that *subjects who used our parallel concordance in the post test made statistically significant improvements over previous translations written with the aid of bilingual dictionaries. (published corpus, emphatics)*

Other bundles appeared in the published writing in the category of emphatics, *It should be noted*, and *It is (also) important to*, mostly served to draw the reader's attention to the limitations of the current or previous research, as in:

It should be noted that *this test is not designed to assess all aspects of collocation knowledge. (published corpus, emphatics)*

It is important to *mention that even though the use of these methods for the teaching of vocabulary and collocations has been widely advocated, there is no specific research on how well they work. (published corpus, emphatics)*

Other bundles used by students in the category of emphatics can be further classified into two sub-categories: *It can be inferred*, *It shows that the* and *It turns out that* performed the same role with the shared bundles, indicating the writer's conclusion; *It is hoped that* and *It is expected that* drew the reader's attention to the implications of the research. The following examples illustrate their different functions:

It can be inferred *from this result that in English teaching, teachers should attract great importance to the improvement of learners receptive collocation competence. (student corpus, emphatics-conclusion)*

It shows that the *learner's mother tongue has a great influence on the appropriateness of the learner's verb-noun collocations. (student corpus, emphatics-conclusion)*

It turns out that *the incorrect use of prepositions constitutes a major problem to most learners at almost every period of English learning. (student corpus, emphatics-conclusion)*

It is hoped that *the findings of the study will shed some light on pedagogical approaches of productive vocabulary acquisition. (student corpus, emphatics-implication)*

It is expected that *the analysis is generalizable for the whole population of Chinese college English learners in their word collocation. (student corpus, emphatics-implication)*

The other bundles in the published corpus were used as hedges (*It is possible that*) or attitude markers (*It is difficult to*, *It is interesting to*, and *It is also worth*), as in:

It is possible that *some of the problematic usage of linking adverbials by apprentice and NNS writers may not simply be a question of under- or over-use, but may also reflect a lack of knowledge of the specific patterns in which a given adverbial typically occurs. (published corpus, hedge)*

It is difficult to *know how far we can generalize these results to other L2 learners of a similar ability. (published corpus, attitude marker)*

However, hedge and attitude marker bundles were absent in the student texts. For the absence of hedge bundles in the student writing, Yang (2013) suggests two reasons typical to Chinese writers: unfamiliarity with the hedge devices and different beliefs in Chinese academic discourse: "the researchers should be authoritative and their results should be as rigorous as possible" (p. 30).

All the other bundles in the student corpus performed the function of attribution (*It is suggested that*, *It is known that*, and *It is believed that*). In line with Hewing and Hewing's (2002) findings,

these attribution bundles were mostly used as general attribution (15 out of 18), with no referencing. The following is an example:

> ***It is suggested that*** *the students should have more practice in writing, for example, keeping diaries, writing book reports, making comments on hot issues, etc. (student corpus, attribution)*

3.4.2. Noun + verb bundles

The student *noun + verb* bundles were featured by *it*-clauses and *it* was used as a reiteration strategy, referring back to the preceding lexical item or sentence as a cohesive element (e.g. *It is completely a*, *It focuses on the*, *It can also be*, *It can be used*, *It is used to*, and *It can reveal not*). The following are two examples:

> *The mutual information score or mutual information index gives a measure of the strength of association between two words.* ***It focuses on the*** *likelihood of two words appearing together within a particular span of words. (student corpus, reiteration it)*
>
> *There are certain classes of English word combinations that cannot be explained with existing syntactic or semantic theories.* ***It is completely a*** *matter of convention. (student corpus, reiteration it)*

In contrast, only one *it*-bundles (*It may be that*) appeared in the published corpus with fairly low frequency. Halliday and Hasan (1976) placed all reiteration forms on a cline from the most specific to the most general: the repetition of the same lexical item, the use of a synonym, near-synonym, superordinate, general noun and pronoun *it*. In comparison to a noun or noun phrase, the use of *it* as a vague reference item in the student writing resulted in a much looser structure.

3.4.3. Conjunction bundles

Table 6 shows the conjunction bundles in the two corpora. The students used more sentence transitions to connect their ideas (e.g. *But*, *So*, *And*). The use of these transitions, however, does not necessarily guarantee the flow of the texts (Hinkel, 2004). On the other hand, the high reliance on sentence transitions might be the result of the training received from their writing courses and

it also indicates that the students may have little awareness or knowledge of alternative cohesive devices.

Table 6: Conjunction bundles

Student bundles	Frequency	Published bundles	Frequency
As is shown in	13	**As can be seen**	23
As can be seen	12	If we look at	10
But it is not	7	As shown in the	8
So when the required	7		
When it comes to	7		
And at the same	6		
As has been pointed	6		
As we all know	6		

3.5. Typical bundles in Chinese student writing

Besides the differences between the student and published bundles discussed above, it was also found that the Chinese students preferred to use a few typical bundles, as shown in Table 7.

Table 7: Typical Bundles in Chinese Student Writing

Typical bundles	Frequency
That is to say	54
With the development of	20
As far as the (…… is concerned)	14
As a matter of (fact)	10
To put it in (another way/other words)	10
Last but not least	10
With the popularization of	8
When it comes to	7
One thing to be (pointed out is that)	6
As we all know	6

Brown (2007) suggests four major sources of errors: interlingual transfer, intralingual transfer, context of learning and communication strategies. Interlingual transfer refers to the interferences from the acquired language (e.g. L1) to the target language. Intralingual transfer places emphasis on the overgeneralization of the rules in the target language. Context of learning includes the negative influences from teachers or materials. Communication strategies are defined as the strategies used by learners to overcome their incompetent language during communication. On the

basis of Brown's (2007) division, the sources of these typical bundles could possibly be attributable to the above four aspects. The use of the bundles *That is to say*, *With the development/popularization of*, *One thing to be (pointed out is that)* and *As we all know,* is likely to be the result of negative transfer from written Chinese *换句话说*, *随着……的发展*, *需要指出的是*, and *众所周知*. The bundle *To put it in (another way/other words)* is possibly the overgeneralisation of the English combination *In other words*. The overuse of the bundles containing fixed expressions, *As a matter of (fact)*, and *Last but not least,* may largely reflect the negative influence from the English pedagogy in mainland China: excessive emphasis placed on English idioms. The bundles *As far as the (…… is concerned)* and *When it comes to* have the equivalence in the published corpus, *In terms of the*, and the highly marked expressions may be consciously used by the students as a replacement. The following examples are these bundles in the student and published texts:

> ***When it comes to*** *the adverb-adjective (Adv-Adj) collocation, few errors are found. (student corpus, typical bundle)*
> ***As far as the*** *verb-noun collocation is concerned, usually there are two kinds of avoidance. (student corpus, typical bundle)*
> ***In terms of the*** *occurrence of referential bundles, it was found that they are more common in conversation than academic prose in both Korean and Spanish, which differs from English lexical bundles. (expert corpus, typical bundle)*

Ellis (1994) argues that defining the sources of errors is largely subject to the biases of researchers and it is impossible to give accurate explanations about errors. Although only the possible sources of these typical bundles can be identified, the analyses will have clear implications for pedagogy.

4. Pedagogical Implications

The present study focuses on a list of sentence initial bundles retrieved from two corpora: a Chinese Master's L2 thesis corpus and a structure-correlated published journal article corpus. The Chinese L2 students in this study were seemingly not so competent in using NP-based bundles. Their NP-based bundles contained more vague nouns, and fewer shell nouns and demonstrative determiners. In comparison to NP-based bundles, the Chinese students were fairly competent in employing PP-

based bundles. The training they received in the writing courses and the transparency of many PP-based bundles (e.g. *On the basis of* and *On the other hand,*) may contribute to the successful acquisition. In contrast, the Chinese students did not use two less transparent ones (*In terms of the* and *In light of the*) although both were found popular in journal articles. VP-based bundles, mainly *In order to* and *to* bundles, featured the sentence starters of the Chinese student writing. Clause-based bundles including anticipatory-*it* bundles, *noun + verb* bundles, and conjunction bundles were also used differently in the two corpora. In comparison to the bundles of the published corpus, the student bundle *It can be seen that* served as a multi-purpose expression and no anticipatory-*it* bundle was used to indicate research limitations, hedge conclusions or express personal attitudes. The *noun + verb* bundles were predominantly composed of the pronoun *it*, loosely linking to the previous text. The conjunction bundles with various sentence transitions were also heavily used in the student writing corpus. Both the use of pronoun *it* and sentence transition bundles partially reflect the students' comparatively limited knowledge of cohesive devices.

These findings have clear implications for EAP writing pedagogy. First, the evidence from our corpus-based comparison suggests the importance of introducing lexical bundles rather than single words to student writers, as lexical bundles (e.g. *It should be noted*) always contain lexico-grammatical patterns (e.g. anticipatory-*it* pattern) and serve certain metadiscourse functions (e.g. emphatics). According to Nation (2013), knowing a word involves knowing its form, meaning and use, and the knowledge of lexical bundles tells student writers where, when and how to use a word. Second, the results of the comparison revealed the comparatively limited writing strategies of Chinese students. For example, besides conjunctions, the advanced L2 students seldom chose other cohesive devices such as shell nouns and demonstrative determiners. Both have been regarded as effective cohesive devices (Flowerdew, 2003; Gray, 2010). EAP teachers can refer to Aktas and Cortes's (2008) work on shell nouns and Gray's (2010) work on demonstratives as examples to demonstrate their lexico-grammatical patterns and discourse functions to students. Third, our study indicates the possible transfer of L1 in Chinese student writing. The use of sentence initial *In order to* and *to* bundles could possibly be the negative transfer of the Chinese phrase为了, literally translated as *in order to*, which usually occurs at the beginning of sentences. The students were probably uncertain of the difference between the English and the Chinese phrase and might have unconsciously transferred the position of their L1 phrase to the target language. As Paquot (2013)

suggests, “EFL teaching needs to counter the default and sometimes misleading L1-related primings in EFL learners’ mental lexicons” (p. 411).

As to the design of bundle learning activities, teachers can refer to Nation’s (2013) three cognitive conditions for vocabulary learning: noticing, retrieval and creative use. Noticing here means seeing a bundle as a learning target and paying attention to it. During reading, teachers can ask students to collect high-frequent sentence starters or sentence initial bundles if a corpus-based tool (e.g. FLAX) is available. Discussions can be organized on the functions of these bundles in writing (e.g. linking function) or on the differences between students’ source and target language in terms of sentence initial bundles. During writing, teachers can use reformulation (Cohen, 1983) strategy to rewrite students’ sentences, preserving their ideas but replacing the inappropriate sentence starters with target bundles. Bundle noticing can be enhanced by comparing the reformulated writing with the original one. Retrieval refers to the recall process of any previously met bundle and creative use occurs when a previously met bundle is used in a new context. Nation (2013) regards creative use as the most effective condition for vocabulary learning. This has also been supported by Peters and Pauwels’s (2015) study on the effect of formulaic sequence instruction: their use of cued output activities, combined both retrieval and creative use stages, turns out to be a more effective approach than recognition activities. With regard to corpus-based language learning approach, Wu, Franken and Witten’s (2010) argument on collocation learning could be transferrable to bundle learning. Noticing can be enhanced with typographically highlighted bundles in texts. Retrieval can be achieved when student writers negotiate the use of an unfamiliar bundle through searching its content word and browsing its multiple contexts. Creative use occurs when students deploy the target bundle in their own writing.

5. Conclusion

In this study, we focused on analysing a list of sentence initial bundles retrieved from two self-built corpora: a Chinese Master’s L2 thesis corpus and a published journal article corpus. In accordance with the practice of many bundle studies, the published articles were used as a good model of academic writing to reveal the divergence of learner bundle production. The focus on sentence initial bundles avoids bundle overlaps and reveals the particular structures and functions of sentence starters. Caution should be taken in attempting to generalise the results to other domains, disciplines, genres or registers. However, the sizes of both corpora are sufficient to

generate salient differences and similarities in the sentence initial bundles between advanced Chinese students' and published writing. The present study suggests some of the possible reasons for the Chinese students' bundle selection and provides advice for improving it. Future research can be designed to further explore the reasons for these choices. One approach could be to move beyond corpus study to involve actual writers in interviews about their choices (e.g. Li, Franken, & Wu, in press). This would generate more complex and nuanced understandings of writers' bundle knowledge (e.g. NP-bundles and anticipatory-*it* bundles) and would thus also better inform ESP writing pedagogy.

References

Ädel, 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*(2), 81-92. doi:10.1016/j.esp.2011.08.004

Aktas, R. N., & Cortes, V. (2008). Shell nouns as cohesive devices in published and ESL student writing. *Journal of English for Academic Purposes, 7*(1), 3-14. doi:10.1016/j.jeap.2008.02.002

Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education, 1*, 105-127.

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers*. Philadelphia, PA: 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. doi:10.1016/j.esp.2006.08.003

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: an initial taxonomy. In G. N. Leech, T. McEnery, A. Wilson, & P. Rayson (Eds.), *Corpus linguistics by the lune*. New York, NY: Peter Lang.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London, United Kingdom: Longman.

Brown, H. D. (2007). *Principles of language learning and teaching* (5th ed.). White Plains, NY: Longman.

Chen, L. (2010). An investigation of lexical bundles in ESP textbooks and electrical engineering introductory textbooks. In D. Wood (Ed.), *Perspectives on formulaic language:*

Acquisition and communication (pp. 107-125). London, United Kingdom: Continuum International Publishing Group.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49.

Cohen, A. D. (1983). Reformulating second-language compositions: A potential source of input for the learner. from ERIC database (ED 228866)

Cortes, V. (2002). Lexical bundles in freshman composition. In R. Reppen, S. M. Fitzmaurice, & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131-145). Amsterdam, Netherlands: 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. doi:10.1016/j.esp.2003.12.001

Cortes, V. (2013). *The purpose of this study is to*: Connecting lexical bundles and moves in research article introductions. *Journal of English for Academic Purposes, 12*(1), 33-43. doi:10.1016/j.jeap.2012.11.002

Crawford, W. J. (2008). Place and time adverbials in native and non-native English student writing. In A. Ädel & R. Reppen (Eds.), *Cropora and discourse: The challenges of different settings* (pp. 267-288). Amsterdam, the Netherlands: John Benjamins Publishing Company.

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

Flowerdew, J. (2003). Signalling nouns in discourse. *English for Specific Purposes, 22*(4), 329-346. doi:10.1016/S0889-4906(02)00017-0

Flowerdew, J. (2013). *Discourse in English language education*. London, United Kingdom: Routledge.

Francis, G. (1986). *Anaphoric nouns*. Birmingham, United Kingdom: English Language Research, University of Birmingham.

Gray, B. (2010). On the use of demonstrative pronouns and determiners as cohesive devices: A focus on sentence-initial this/these in academic prose. *Journal of English for Academic Purposes, 9*, 167-183.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London, United Kingdom: Longman.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar* (3rd ed.). London, United Kingdom: Arnold.

Herbel-Eisenmann, B., & Wagner, D. (2010). Appraising lexical bundles in mathematics classroom discourse: Obligation and choice. *Educational Studies in Mathematics, 75*(1), 43-63.

Hewings, M., & Hewings, A. (2002). "It is interesting to note that ...": A comparative study of anticipatory 'it' in student and published writing. *English for Specific Purposes, 21*, 367-383.

Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning, 12*, 111-132.

Hinkel, E. (2002). *Second language writer's text: Linguistic and rhetorical features*. NJ: Lawrence Erlbaum Associates.

Hinkel, E. (2004). *Teaching academic ESL writing: Practical techniques in vocabulary and grammar*. Mahwah, NJ: Lawrence Erlbaum Associates.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London, United Kingdom: Continuum.

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

Hyland, K. (2008b). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21. doi:10.1016/j.esp.2007.06.001

Hyland, K. (2008c). Persuasion, interaction and the construction of knowledge: Representing self and others in research writing. *International Journal of English Studies, 8*(2), 1-23.

Ivanič, R. (1991). Nouns in search of a context: A study of nouns with both open- and closed-system characteristics. *International Review of Applied Linguistics, 29*, 93-114. doi:10.1515/iral.1991.29.2.93

Jablonkai, R. (2010). English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes, 29*(4), 253-267. doi:10.1016/j.esp.2010.04.006

Jiang, F., & Hyland, K. (2015). 'The fact that': Stance nouns in disciplinary writing. *Discourse Studies*, 1-22. doi:10.1177/1461445615590719

Jiang, F., & Hyland, K. (2016). Nouns and academic interactions: A neglected feature of metadiscourse. *Applied Linguistics*, 1-25. doi:10.1093/applin/amw023

Jiang, F., & Hyland, K. (2017). Metadiscursive nouns: Interaction and cohesion in abstract moves. *English for Specific Purposes, 46*, 1-14. doi:10.1016/j.esp.2016.11.001

Kaneyasu, M. (2012). *From frequency to formulaicity: Morphemic bundles and semi-fixed constructions in Japanese spoken discourse.* (Doctoral dissertation), University of California, Los Angeles, CA. Retrieved from ttps://escholarship.org/uc/item/1zp613xj#page-1

Kim, Y. (2009). Korean lexical bundles in conversation and academic texts. *Corpora, 4*(2), 135-165. doi:10.3366/E1749503209000288

Li, L. (2016). Sentence initial bundles in L2 thesis writing: A comparative study of Chinese L2 and New Zealand L1 postgraduates' writing (Thesis, Doctor of Philosophy (PhD)). University of Waikato, Hamilton, New Zealand. Retrieved from https://hdl.handle.net/10289/10862

Li, L., Franken, M., & Wu S. (in press). Chinese postgraduates' explanation of the sources of sentence initial bundles in their thesis writing. RELC Journal.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge, United Kingdom: Cambridge University Press.

Neely, E., & Cortes, V. (2009). A little bit about: Analyzing and teaching lexical bundles in academic lectures. *Language Value, 1*(1), 17-38.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics, 11*(3), 283-283. doi:10.1075/ijcl.11.3.04nes

Pan, F., Reppen, 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. doi:10.1016/j.jeap.2015.11.003

Pang, P. (2009). A study on the use of four-word lexical bundles in argumentative essays by Chinese English-majors: A comparative study based on WECCL and LOCNESS. *CELEA Journal, 32*(3), 25-45.

Paquot, M. (2013). Lexical bundles and L1 transfer effects. *International Journal of Corpus Linguistics, 18*(3), 391-417. doi:10.1075/ijcl.18.3.06paq

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191-226). London, United Kingdom: Longman.

Peters, E., & Pauwels, P. (2015). Learning academic formulaic sequences. *Journal of English for Academic Purposes, 20*, 28-39. doi:10.1016/j.jeap.2015.04.002

Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System, 42*(1), 220-231. doi:10.1016/j.system.2013.12.003

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London, United Kingdom: Longman.

Schnur, E. (2014). Phraseological signaling of discourse organization in academic lectures: A comparison of lexical bundles in authentic lectures and EAP listening materials. *Yearbook of Phraseology, 5*(1), 95-122. doi:10.1515/phras-2014-0005

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford, United Kingdom: Oxford University Press.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*(3), 214-225. doi:10.1016/j.jeap.2013.05.002

Tracy-Ventura, N., Cortes, V., & Biber, D. (2007). Lexical bundles in speech and writing. In G. Parodi (Ed.), *Working with Spanish corpora* (pp. 217-231). London, United Kingdom: Continuum International Publishing.

Tse, P., & Hyland, K. (2009). Academic lexis and disciplinary practice: Corpus evidence for specificity. *International Journal of English Studies, 9*(2), 111-129. doi:10.6018/ijes.9.2.90781

Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal, 42*(2), 155-166. doi:10.1177/0033688211407295

Williams, J. M. (2003). *Style: Ten lessons in clarity and grace* (7th ed.). New York, NY: Addison-Wesley.

Wu, S., Franken, M., & Witten, I. H. (2009). Refining the use of the web (and web search) as a language teaching and learning resource. *Computer Assisted Language Learning, 22*(3), 249-268.

Wu, S., Franken, M., & Witten, I. H. (2010). Supporting collocation learning with a digital library. *Computer Assisted Language Learning, 23*(1), 87-110.

Xu, F. (2012). 中国学习者英语学术词块的使用及发展特征研究 [The use and developmental features of lexical bundles in Chinese learners' English academic writing]. *外语与外语教学 [Foreign Languages and Their Teaching], 4*, 51-56. doi:10.13564/j.cnki.issn.1672-9382.2012.04.013

Yang, Y. (2013). Exploring linguistic and cultural variations in the use of hedges in English and Chinese scientific discourse. *Journal of Pragmatics, 50*(1), 23-36. doi:10.1016/j.pragma.2013.01.008

Appendix: Sentence initial bundles in frequency order

Student bundles	Tokens	Published bundles	Tokens
That is to say	45	**On the other hand**	36
In the present study	43	**On the basis of**	15
On the other hand	32	**In the case of**	13
On the basis of	24	**At the same time**	12
With the help of	19	**In the present study**	11
In the process of	19	It is important to	10
At the same time	18	**The results showed that**	10
With the development of	17	**As can be seen**	12
It can be seen	17	**It was found that**	8
In order to make	15	It should be noted	8
It was found that	14	**On the one hand**	8
It is found that	14	The fact that the	8
As far as the	12	In terms of the	7
It is obvious that	11	**It is clear that**	7
The present study is	11	There was no significant	6
The results showed that	11	**With the help of**	6
As is shown in	11	It is possible that	6
In the case of	10	It is difficult to	6
On the one hand	10	It is interesting to	6
As can be seen	10	**The results indicated that**	5
The results of the	9	As a result of	5
The second type of	9	If we look at	5
It is hoped that	9	**The results of the**	5
The following table shows	9	The results of this	5
One of the most	8	Table 1 shows the	5
As a matter of	8	This leads us to	5
It is also found	8	Table 2 shows that	5
Based on the above	8	**With regard to the**	5
To put it in	8	The analysis of the	4
The reason might be	8	For the purposes of	4
We need a better	8	Table 2 presents the	4

The results indicate that	8	In this article we	4
The results show that	8	**In addition to the**	4
Last but not least,	8	As shown in the	4
With the popularization of	7	The aim of this	4
In view of the	7	The findings of this	4
In addition to the	7	One possible explanation for	4
It is suggested that	7	It may be that	4
It is completely a	7	The first step in	4
Today it is very	7	It is also worth	4
Twenty years ago it	7	There is also a	3
The information such as	6	Figure 4 shows the	3
In spite of the	6	There was also a	3
In the same way	6	For the purpose of	3
In the following section,	6	In the majority of	3
It can be inferred	6	Table 3 shows the	3
It is known that	6	With the exception of	3
It is expected that	6	This is an important	3
It is clear that	6	The study shows that	3
This result indicates that	6	The implication is that	3
The present study adopts	6	Table 2 shows the	3
The result showed that	6	Figure 3 shows that	3
The third chapter presents	6	The first of these	3
It is based on	6	The findings of the	3
But it is not	6	In light of the	3
So when the required	6	Despite the fact that	3
When it comes to	6	In addition to these	3
Ten years ago it	6	These results show that	3
The examples of the	5	This is because the	3
The range of the	5	One of the reasons	3
The size of the	5	The purpose of this	3
The total number of	5	**The size of the**	3
One thing to be	5	**The total number of**	3
In one of his	5	The design of the	3
In the following part	5	The use of these	3

With regard to the	5	In their study of	3
It shows that the	5	In the first part	3
It is believed that	5	**It is obvious that**	3
It turns out that	5	It is also important	3
Based on this general	5		
In order to get	5		
In order to have	5		
It focuses on the	5		
The following examples are	5		
Another example is the	5		
It can also be	5		
It can be used	5		
The same is true	5		
The reason may be	5		
The samples are all	5		
Each component may be	5		
The results revealed that	5		
This study focuses on	5		
The result shows that	5		
It is used to	5		
It can reveal not	5		
This shows that what	5		
And at the same	5		
As has been pointed	5		
As we all know	5		
There are two possible	5		