The Different Use of Lexical Bundles by L1 Chinese Master Students and L1 English Experts Based on Corpus

Li Xin*
School of Foreign Languages and Literature, Shandong University, Jinan, Shandong, China
**Corresponding Author.*

Abstract: Lexical bundles have been proven essential for non-native students to develop their competence. However, the comparison of lexical bundles between L2 students' thesis writing and L1 native published research articles have not undergone in-depth analysis. Furthermore, it is critical to draw on corpus to increase master students' proficiency in academic writing in English. Drawing on a 1.16-million-word self-built corpus of academic lectures in the field of linguistics, we tend to compare the use of lexical bundles in terms of Biber's structural and Hyland's functional taxonomy. The findings indicate that in general, the number of types of four-word lexical bundles in L1 English experts' published articles are more than that in L1 Chinese. This research may help to improve the fluency, variety and diversity of English discussion and conclusion part of master students.

Key words: Lexical Bundles; L2 Masters' Theses; L1 Experts' Published Articles

1. Introduction

Lexical bundles, the high-frequency word combinations in natural discourse, have been proven to be essential for non-native students to develop their competence. Previous studies have focused on the comparison of lexical bundles in terms of different disciplines, different genres and different contexts. However, the comparison of lexical bundles between L2 students' thesis writing and L1 native published research articles have not undergone in-depth analysis. Furthermore, it is critical to draw on corpus to increase master students' proficiency in academic writing in English as writing in foreign language is a core requirement for English-major students and it is hard to develop in a short period of time if correct instructions are ignored.

Our study aims to help masters to use lexical bundles in a more appropriate and meaningful way and enhance their thesis writings through exploring the preferred lexical bundles used by expertise. But as everyone is not born with academic writing skills that are non-native for everyone [1], our intention is not to regard published academic writing as a rule of indestructibility but to bring some insight into the different use of lexical bundles between L1 expert writing and L2 masters' writing. In this way, masters could be guided about the lexical bundles in academic writing.

2. Literature Review

2.1 Formulaic Sequences

Formulaic sequences, also known as recurrent combinations of words, seem to be prefabricated and stored in the mind of the speaker and can be mobilized as a whole instead of relying on grammatical analysis [2]. They are frequently used in discourse and are pervasive in language use [3]. Non-native speakers may be confusing as no specific rules are set for the preferred and frequent word sequences. Thus, sensitivity to the correct word sequences is necessary for non-native speakers to approach native-like competency. In this way, language could be more predictable and understandable to the other party. To avoid verbose or informal expressions, it is feasible to gain good control over formulaic sequences. Without this skill, it's difficult to meet the expected level of proficiency [4].

As a cover term for lexical bundles, formulaic sequences are essential for developing native-like proficiency in academic writing and are grown in popularity in English-teaching due to its formulaic nature of pragmatics norms [5]. Formulaic sequences prepare the way for writers to formulate their research art because the writer is using expressions that pre-exist in their cognitive repertoire instead of making

every sentence from scratch [6]. Thus, specific instruction strategies such as repetition could be implemented to improve students' knowledge about formulaic sequences.

2.2 Lexical Bundles

Lexical bundles are word sequences that exhibit frequent co-occurrence in natural discourse and serve as crucial "building block of discourse" [7] with significant implications for effective communication. They have taken different terms such as clusters, lexical phrase, n-grams and multi-word expressions [8, 9]. Different opinions occur regarding the feature of lexical bundles. Idiomaticity, fixedness and other qualities relevant to word combinations are emphasized by Cortes [10] to describe lexical bundles while Biber [11] reveals the non-idiomatic and non-complete nature of lexical bundles.

Lexical bundles have been explored in terms of different registers including classroom discourse, university teaching and textbooks, proficiency exams and academic writing [12]. the difference between spoken and written register is particularly evident. For example, Biber and Barbieri [7] show that lexical bundles appear frequently in written course management and are often used to organize discourse and express their stance in spoken register. Previous studies also reveal the difference between conversation and academic writing in terms of structure and function. In structural taxonomy, academic writing is more phrasal while conversation is more clausal. In functional taxonomy, academic writing are realized through heavy use of referential registers while conversation relies more on stance and discourse organisers.

This study aims to offer a deeper understanding for L2 master students concerning the use of lexical bundles by comparing their thesis writings with L1 experts' published articles. This study is guided by the following questions: What similarities and disparities are discernible in the structural types of lexical bundles used by L1 Chinese masters and L1 English experts?; What similarities and disparities are discernible in the functional types of lexical bundles used by L1 Chinese masters and L1 English experts?; If differences exist, what are the factors that contribute to them?

3. Method

3.1 Corpora

Data for our study are composed of written texts including L2 masters' thesis and L1 published research articles. One of the sub-corpus, masters' thesis writing corpus (MTWC), is established by selecting masters' thesis in the discipline of linguistics from CNKI and WANFANG DATA from 2014 to 2022.135 thesis are randomly selected with a total number of 579294 words. It should be noted that the masters' thesis collected on the two data platforms are excellent theses, representing a relatively high level of writing among masters' students.

The sub-corpus, experts' published writing corpus (EPWC), are built by selecting articles from three highly-regarded journals with significant influence, with an average impact factor of 3.7, including System, Applied linguistics, and English for Specific Purpose. In order to maintain a balance in word count, a total of 298 articles comprising 588800 words from 2014-2022 are chosen from these three publications. Detailed information on article choice is provided in the following table.

3.2 Lexical Bundle Extraction

Four-word lexical bundles have particularly caught our attention because four-word lexical bundles may incorporate three-word bundles within their framework [13] and provides a clear range of functions and structures [14]. Moreover, the quantity of four-word lexical bundles is frequently within a feasible scope (approximately 100) for manual classification and analysis [15]. Dispersion criteria is necessary to set to circumvent any eccentricities brought by particular writers. the extracted lexical bundles occur at least 3 texts to 5 texts or 10% of texts, depending on the size of sub-corpora. Bundles that meet the frequency value appear in at least 3 texts for 50, 000 sub-corpora, 4 texts for 100, 000 sub-corpora and 5 texts for 200, 000 sub-corpora [7]. Based on this criteria, the dispersion threshold in this study spans at least 5 texts.

We draw on *Antconc* to extract lexical bundles that meet the requirement. Based on the set frequency (25 times per million words) and the total number of tokens in the two sub corpora, four-word lexical bundles that occur at least 14 times across at least 5 texts are selected.

However, not all of these word bundles satisfy our criteria. Lexical bundles that match frequency and dispersion rates are methodically handpicked. After manual screening based on the aforementioned criteria, 142 and 111 four-word lexical bundles that satisfied the norms were extracted in MTWC and EPWC respectively.

3.3 Functional and Structural Taxonomy of Lexical Bundles

Drawing upon Biber et al. 's [12] classification, Hyland [14] has developed a novel functional taxonomy of lexical bundles, primarily divided into three categories: research-oriented bundles (organization of writers' activities and real-world experience); text-oriented bundles (structuring the text and its intended message or point of contention); participant-oriented bundles (expressing writers' opinions or attitudes or speaking directly to readers). the structural analysis of lexical bundles adopted by this study is based on Biber et al. 's [16] structural taxonomy. Three main categories include NP-based bundles, PP-based bundles and VP-based bundles. NP-based and PP-based bundles are phrasal bundles while VP-based bundles are clausal bundles.

4. Results and Discussion

4.1 General Comparison of Lexical Bundles

More lexical bundle types (142) are seen in non-native students' theses than in native professionals' articles (111). Students' increased dependence on lexical bundles is consistent with the prior research by Hyland [14], which implies that students are less confident or adept in their thesis. Another factor should be mentioned that due to the nature and goal of an MA's thesis, students tend to demonstrate their expert-like proficiency and validate their capacity to meet graduation criteria.

4.2 Distributions of Structural Taxonomies of Lexical Bundles

Different distributions of structural taxonomies of bundles in MTWC and EPWC are presented in **Table 1**. EPWC has a higher type/token ratio in each type of bundle than MTWC, implying that L1 experts have a more abundant vocabulary than L2 masters.

For structural taxonomy, PP-based bundles account for the largest part in both corpora, but the share of NP-based bundles are lower in L1 Chinese students than that in L1 English experts. It demonstrates students' increasing awareness of the use of phrasal bundles but existent ignorance of NP-based bundles.

The highest share of subcategories in non-native masters' thesis are NP with of-phrase fragment and PP with embedded of-phrase (accounting for the same proportion-19%), which indicates increasing awareness of the use of phrasal bundles and their higher proficiency level among L2 learners for the nature of sample we extract (the articles collected by the CNKI are all excellent masters' thesis in linguistics). However, the total types of NP-based bundles are still the smallest in masters' thesis and there is a significant difference in the use of NP-based bundles between non-native MA's thesis writing and native professional articles.

The relatively high proportion of VP-bundles in both L1 Chinese and L1 English writings can be explained by the nature of soft science and the extract of discussion and conclusion part, focusing on the clarification of relationships based on the previous findings and pointing out the contributions, limitations or future directions of the current study.

4.2.1 NP-based Bundles

Although MTWC has more types than EPWC in terms of NP with of-phrase fragment, the proportion of lexical bundles in EPWC is relatively higher. There are 9 bundles in NP with of-phrase fragment shared by both MTWC and EPWC. Despite the fact that both sub-corpora use these lexical bundles, there are some substantial use variances. To emphasize findings of the current study, L1 experts tend to select the demonstrative pronoun "this" rather than the definite article "the". Moreover, in terms of NP with of-phrase fragment, MTWC offers more quantity-related bundles that are not shared by EPWC. L1 experts draw on these bundles for illustrating the properties and functions of something or explaining the connotation of something and its connection with previous research. This finding is consistent with previous studies by Chen and Baker [15] and Shaojie Zhang et al. [17] and it may be due to masters' lack of knowledge and proficiency in using this kind of lexical bundles.

Table 1 Different Distributions of Structural Taxonomies of Bundles in Subcorpora

*=significant p < 0.05; **=significant p < 0.01

Categories	Subcategories	Type		Token		Type/Token ratio	
		MTWC	EPWC	MTWC	EPWC	MTWC	EPWC
NP-based bundles	NP with of-phrase fragments**	19.0% (27)	21.6% (24)	16.8% (3164)	21.8% (2720)		
	NP with other postmodifier fragments**	4.2% (6)	7.2% (8)	3.0% (568)	7.1% (892)		
	Other NPs**	3.5% (5)	1.8% (2)	2.9% (552)	1.5% (188)		
	Total**	26.7% (38)	30.6% (34)	22.7% (4284)	30.4% (3800)	0.00887	0.00894
PP-based bundles	PP with embedded of-phrase**	19.0% (27)	19.8% (22)	20.7% (3884)	21.3% (2660)		
	Other PP fragments**	17.6% (25)	13.5% (15)	22.2% (4180)	19.7%% (2460)		
	Total**	36.6% (52)	33.3% (37)	42.9% (8064)	41.0% (5120)	0.00645	0.00722
VP-based bundles	Anticipatory it +VP/adj phrase**	9.2% (13)	8.1% (9)	10.9%(2044)	8.1% (1012)		
	Passive verb + PP**	2.8% (4)	1.8% (2)	1.6% (308)	1.5% (188)		
	Copula be +NP/adj phrase**	2.1% (3)	0% (0)	1.6% (292)	0% (0)		
	(VP) + that-clause fragment**	5.6% (8)	0.9% (1)	4.4% (820)	0.5% (64)		
	(Verb/adjective) to-clause fragment**	4.2% (6)	9.0% (10)	5.2% (960)	6.0% (752)		
	Pronoun/NP + be fragment	4.9% (7)	7.2% (8)	3.2% (600)	4.8% (604)		
	Adverbial clause fragment**	2.1% (3)	1.8% (2)	2.5% (476)	1.6% (196)		
	VP with active verb**	0% (0)	1.8% (2)	0% (0)	1.1% (136)		
	Total**	31.0% (44)	30.6% (34)	29.3% (5500)	23.6% (2952)	0.00800	0.01151
Others		5.6% (8)	5.4% (6)	5.0% (940)	5.0% (624)	0.00851	0.00992

4.2.2 PP-based Bundles

MTWC and EPWC have roughly the same proportion in terms of PP with embedded of-phrase, but the types and tokens in MTEC are more than those in EPWC. 12 bundles are shared by both sub-corpora and the majority of them are presented as "in/on/at+the+n. +of". MTWC and EPWC incorporate the same lexical bundle that may play a different role in the specific context. For example, L2 masters often use “at the end of” to denote the location where something occurs, while L1 experts usually use it to indicate the role that something plays in this position.

4.2.3 VP-based Bundles

MTWC embraces more types and a higher proportion than EPWC in terms of anticipatory it +VP/adj phrase. And in MTWC, this type of lexical bundle accounts for the most in the subcategories of verb-based bundles. Only 2 bundles are shared by MTWC and EPWC, including “it is important to” and “it was found that”. Apart from “it was found that”, L2 masters also heavily uses “it is found that” (101 times in MTWC) to report results or findings, which is uncommon in experts’ published articles. It indicates masters’ mixed use of tenses due to their non-proficiency and unfamiliarity with the specific use of tenses in academic register. For example, in (1):

(1) Generally, *it was found that* with higher Chinese language proficiency, longer terminable TC-units consisting of more dependent single TC-units were produced. (EPWC)

The greater number of "that-clausal" bundles

in masters' thesis demonstrates the colloquial feature in their writings. the use of bundles by L2 writers seem to be fusions that are not totally native-like.

4.3 Distributions of Functional Taxonomies of Lexical Bundles

Different distributions of functional taxonomies of bundles in MTWC and EPWC have been shown in **Table 2**. MTWC has a higher type in the taxonomy of research-oriented bundles, text-oriented bundles and participant-oriented bundles.

Table 2. Different Distributions of Functional Taxonomies of Bundles in Subcorpora

*=significant p < 0.05; **=significant p < 0.01

Categories	Subcategories	Type		Token		Type/Token ratio	
		MTWC	EPWC	MTWC	EPWC	MTWC	EPWC
Research-oriented bundles	Location**	4.2% (6)	3.6% (4)	4.2% (788)	4.6% (576)		
	Procedure**	7.0% (10)	7.2% (8)	7.7% (1440)	8.1% (1012)		
	Quantification**	12.0% (17)	10.8% (12)	9.6% (1804)	8.2% (1028)		
	Description**	11.3% (16)	13.5% (15)	8.4% (1584)	11.7% (1464)		
	Topic*	1.4% (2)	1.8% (2)	3.0% (564)	1.1% (132)		
	Total**	35.9% (51)	36.9% (41)	32.9% (6180)	33.7% (4212)	0.0083	0.0097
Text-oriented bundles	Transition signals**	11.3% (16)	10.8% (12)	12.8% (2404)	12.3% (1532)		
	Resultative signals**	9.9% (14)	9.0% (10)	11.6% (2172)	9.7% (1216)		
	Structuring signals**	15.5% (22)	14.4% (16)	12.4% (2328)	16.2% (2020)		
	Framing signals**	9.9% (14)	14.4% (16)	13.1% (2476)	15.6% (1944)		
	Total**	46.6% (66)	48.6% (54)	49.9% (9380)	53.8% (6712)	0.0070	0.0080
Participant-oriented bundles	Stance features**	10.6% (15)	11.7% (13)	9.1% (1704)	10.1% (1256)		
	Engagement**	7.0% (10)	2.7% (3)	8.1% (1524)	2.5% (316)		
	Total**	17.6% (25)	14.4% (16)	17.2% (3228)	12.6% (1572)	0.0077	0.0102

4.3.1 Research-oriented Bundles

Writers use research-oriented bundles in organizing their real-world experiences and activities. Masters tent to use more location, quantification and description bundles.

In terms of quantification bundles, masters use them to present concrete quantitative information that are often in the form of "the+n. +of" such as "the average value of", "the frequency of the", "the degree of the" and "on the index of". By contrast, experts often use quantitative bundles to convey conceptual information with embedded clause modifiers such as "the degree to which" and "the extent to which", as in (2):

(2) Through the comparison, it can be noticed that *the frequency of the* second personal pronoun "you" is seldom used in both trump and Hillary's election speeches. (MTWC)

4.3.2 Text-oriented Bundles

Text-oriented bundles is prepared to organize the text and its meaning as an argument [30]. Text-oriented bundles account for the largest part of both corpora.

Transition signals are used either to set up additive links between elements or compare

and contrast different elements. the bundles in transition signals shared by students and experts are “on the other hand”, “on the one hand”, “as well as the” and “in line with the”. the non-shared bundles “in a way that” and “the way in which”, as initiators of embedded clauses, used by L1 experts for either comparison - a comparison with other methods or instruments to emphasize the superiority of the former, or for further clarification of the current method, as shown in (3):

(3) Among other advantages, video can highlight non-verbal actions *in a way that* other feedback modalities cannot.

Resultative signals are use to identify inferential or causal relationships between items. What both have in common is "the result of the", which appears the most frequently in resultative signals. Experts prefer to draw on “the result of the” to support their hypothesis and answer.

4.3.3 Participant-oriented Bundles

Participant-oriented bundles are used to address writers and readers in the text [30]. Participant-oriented bundles account for the smallest proportion in both corpora, but EPWC has fewer categories and a lower proportion than MTWC. the majority of participant-oriented bundles are made up of stance bundles. Both non-native masters and native experts tend to use impersonal epistemic bundles to convey their evaluation or attitude. However, stance bundles in non-native students’ writing occur less frequently than those in native experts’ articles, which may be a result of culture influence as Chinese prefer to take a non-interventionist position. In line with Zhang et al. ’s [17] findings, students prefer to express a certain attitude through explicit evaluation to convince readers of the reliability of their findings. This may be due to the fact that students' thesis writing focuses more on demonstrating their expertise in academic research, whereas published articles hedge the discourse to avoid potential threats to their claims and allow plenty of leeway for readers to discuss.

5. Conclusion

This study has discovered different distributions of lexical bundles between L1 masters’ thesis writings and L2 experts’ published articles in their discussion and conclusion section. As authors write different sections for different purposes, this study has further refined the disparities in academic writings between native speakers and non-native speakers by highlighting the common components of academic writings - discussion and conclusion part. Through detailed analysis from the structural and functional taxonomy, it is found that there is still a large gap between L1 masters’ thesis writings and L2 experts’ published articles in their use of lexical bundles.

Pedagogically, it is worth integrating the preferred use of lexical bundles into their teachings to shed light on L2 masters’ writings. Explicit instructions should be given to improve masters’ knowledge about lexical bundles by comparing different distributions of lexical bundles in terms of structural and functional taxonomy. It is important to help them notice frequent occurring lexical bundles and raise awareness of using them through repeated exposure and classroom activities [14]. Due to the influence of multiple factors such as different paradigms, there is no universal glossary for lexical bundles in academic writings. Therefore, it is critical to guide students in grasping the qualities and essence of lexical bundles. Moreover, it would be helpful to draw support from *AntConc* to deepen masters’ impression by presenting the frequency and collocation of lexical bundles.

However, there are still some limitations about this study. Due to the relatively small corpus of our study, future research can form a larger corpus for more accurate comparisons and collect materials across disciplines to explore their differences. Moreover, five- or six-word lexical bundles could be also considered in terms of structural and functional taxonomy. Furthermore, the use of lexical bundles in different sections and their locations in texts could also be compared based on their different moves.

References

[1] Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. Journal of English for Academic Purposes, 30, 38-52.

[2] Wray, A. (1999). Formulaic language in learners and native speakers. Language Teaching, 32(4), 213-231.

[3] Nattinger, J. R., & DeCarrico, J. S. (1992).

Lexical Phrases and Language Teaching. Oxford University Press.
[4] Cowie, A. P. (1992). Multiword lexical units and communicative language teaching. Vocabulary and Applied Linguistics, 1-12.
[5] Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and lexical phrases. Phraseology: Theory, Analysis and Applications, 145-160.
[6] Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. Journal of Second Language Writing, 16(3), 129-147.
[7] Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. English for Specific Purposes, 26(3), 263-286.
[8] Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid. Formulaic Sequences: Acquisition, Processing and Use, 127-151.
[9] Nattinger, J. R., & DeCarrico, J. S. (1992). Lexical Phrases and Language Teaching. Oxford University Press.
[10] Cortes, V. (2002). Lexical Bundles in Academic Writing in History and Biology. Northern Arizona University.
[11] Biber, D. (2006). Stance in spoken and written university registers. Journal of English for Academic Purposes, 5(2), 97-116.
[12] Biber, D., Conrad, S., & Cortes, V. (2004). If you look at.: Lexical Bundles in University Teaching and Textbooks. Applied Linguistics, 25(3), 371–405.
[13] Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. English for Specific Purposes, 23(4), 397-423.
[14] Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. English for Specific Purposes, 27(1), 4-21.
[15] Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. Language, Learning and Technology, 14(2), 30-49.
[16] Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). Longman Grammar of Spoken and Written English. Pearson Education: Harlow, UK.
[17] Zhang, S., Yu, H., & Zhang, L. J. (2021). Understanding the Sustainable Growth of EFL Students' Writing Skills: Differences between Novice and Expert Writers in Their Use of Lexical Bundles in Academic Writing. Sustainability, 13(10), 1-17.
[18] Bychkovska, T., & Lee, J. J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. Journal of English for Academic Purposes, 30, 38-52.

