The European Proceedings of
Social & Behavioural Sciences
EpSBS

ISSN: 2357-1330

https://doi.org/10.15405/epsbs.2019.09.46

INCoH 2017
The Second International Conference on Humanities

A CORPUS ANALYSIS OF PHRASEOLOGICAL SEQUENCES IN ACADEMIC WRITING

Ang Leng Hong (a)*
*Corresponding author

(a) School of Humanities, Universiti Sains Malaysia, 11800 USM, Malaysia, lenghong@usm.my

Abstract

Findings from corpus-based research have illustrated the importance of phraseology in language acquisition and language use. Many studies which explored the facet of language acquisition and language use have focused on phraseological sequences such as lexical bundles, collocations, lexical phrases and etc. More recent studies have forayed into the use of phraseological sequences in different academic disciplines. The current study aims to add to this growing body of literature by examining a type of phraseological sequence, namely the lexical bundles that are employed by research article writers in the field of International Business Management (IBM). The study is concerned with the identification of lexical bundles. By adopting Mutual Information (MI) statistical measure, the corpus analysis tools, *Collocate 1.0 and AntConc 3.4.1w* were used to identify and extract three- to five-word lexical bundles commonly found in IBM research articles. Results indicate that three-word lexical bundles, i.e., *more likely to*, *in order to*, *as well as*, *in terms of* and *the number of* are most widely found in IBM research articles. This study contributes to the development of English for Academic Purposes (EAP) courses as the lexical bundles identified could be used in EAP settings to develop the relevant teaching materials for the courses. This study also contributes towards raising the awareness of the phraseological nature and tendency in academic writing, particularly in discipline-specific research articles.

Keywords: Phraseological sequences, lexical bundles, academic writing, English for Academic Purposes.

https://doi.org/10.15405/epsbs.2019.09.46
Corresponding Author: Ang Leng Hong
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

1. Introduction

The advances in using computer-mediated research methodology to explore various language features, particularly the phraseological sequences have enabled researchers to continue defining and refining the understanding of various types of phraseological sequences in different genres. Seeing the need for novice academic writers to learn how to write academically and fluently, scholars have begun to look at how words co-occur frequently in academic discourse to form phraseological sequences useful for understanding the meanings and discourse functions in academic texts. In relation to various phraseology research in academic genres, scholars have looked at continuous multi-word sequences such as collocations (Frankenberg-Garcia, 2018; Green & Lambert, 2018; Lei & Liu, 2018), idioms (Hsu, 2014; Liu, 2017), lexical bundles (Adel & Erman, 2012; Qin, 2014; Shin, Cortes, & Yoo, 2018; Wright, 2019). Among the various types of continuous phraseological sequences, lexical bundles have received considerable attention in recent years. This phenomenon was mainly attributed to the seminal work, *The Longman Grammar of Spoken and Written English* by Biber, Johansson, Leech, Conrad, and Finegan (1999) in which they proposed the construct of *lexical bundles* in differentiating between academic prose and conversation. This seminal work deserves attention here as most studies on lexical bundles are largely based on the definition and framework proposed by Biber et al. (1999). Their study of lexical bundles was based on a corpus analysis of multi-million-word language corpora representing academic prose and conversation. Following frequency-based approach, Biber et al. (1999) identified the lexical bundles and compared their structural properties in written and spoken registers.

Recently, there is a growing awareness of the necessity of incorporating explicit teaching of lexical bundles in English for Academic Purposes (EAP) curricula. This is evidenced by the empirically derived lists of phraseological sequences, for instance, the Academic Formulas List (AFL) by Simpson-Vlach and Ellis (2010). This AFL serves as a good start for placing the teaching and learning of phraseological expressions high on the agenda of linguists and language instructors in the field of EAP.

2. Problem Statement

With the flourishing of phraseology research since the last decade, there are a number of phraseological studies on academic writing. A notable one was by Simpson-Vlach and Ellis (2010) that focused on compiling lists of lexical bundles (AFL) common to many disciplines. Nevertheless, there is a need to look at phraseological sequences specific to different disciplines as the learning in EAP classrooms would become more effective if it is based on discipline-specific conventions (Hyland, 2002, 2006). It is indisputable that there are considerable amount of formalities in academic writing that are highly characterised by the use of discipline-specific vocabulary. The issue of specificity is therefore a challenge for language instructors in the field of EAP as they need to be familiar with the vocabulary and phraseological sequences commonly employed by writers in the academic settings in order to facilitate novice writers and learners in their academic learning and writing. Also, it has been discovered that there is a dearth of studies relevant to the field of International Business Management. Given the envisioned pedagogical value of lexical bundles in specific EAP courses, this study thus addresses the issue of specificity in EAP by identifying and compiling list of lexical bundles that are frequently employed by International Business Management research article writers.

https://doi.org/10.15405/epsbs.2019.09.46
Corresponding Author: Ang Leng Hong
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

3. Research Questions

Specifically, this study addresses the following question:

1) What are the most frequent lexical bundles used in the journal articles in International Business Management?

4. Purpose of the Study

Given the importance of phraseological sequences in academic writing, and the increasing prevalence of analyses into language use through corpus-based methods, the present study intends to use corpus-based methods to identify the most frequently used lexical bundles in a corpus consisting of International Business Management (IBM) journal articles.

5. Research Methods

The present study employed corpus-based methods to identify frequently used three- to five- word lexical bundles in a one-million word corpus with 138 original research articles taken from two Thomson Reuters-indexed journals relevant to the field of IBM that achieve satisfactory impact factor yearly.

5.1. Identification of Lexical Bundles

The aim of the study was to identify and the most frequent lexical bundles in IBM corpus. In accordance with Biber et al. (1999), lexical bundle was broadly defined as frequently recurring continuous sequence of words. This study focused on three- to five-word lexical bundles. The corpus analysis tool, *Collocate 1.0* (Barlow, 2004) was used to retrieve lexical bundles automatically by setting the span options as well as the statistics options, i.e., frequency and Mutual Information (MI). Following the literature, the minimum cut-off frequency and MI score were set at 20 times per million words and 3.00 and above, respectively. *Collocate 1.0* extracted a total of 1714 three-word sequences, 270 four-word sequences and 25 five-word sequences. As the extraction was automatic, the list of multi-word sequences extracted by *Collocate 1.0* needed to be manually inspected to exclude the irrelevant and meaningless word combinations. The last step in identifying lexical bundles was to check the dispersions of the multi-word sequences in IBM corpus. The corpus analysis tool, *AntConc 3.4.1w* (Anthony, 2015) was used to check the dispersions of lexical bundles in IBM corpus. Based on the literature, a lexical bundle has to occur in three to five texts (Biber & Barbieri, 2007) or 10% of texts to avoid idiosyncrasies of particular writers (Hyland, 2008). In the present study, it was determined that lexical bundles which occur in at least 10% of the texts were qualified as the lexical bundles for this study.

6. Findings

A total of 1055 lexical bundles of varying lengths were identified in IBM corpus. The lexical bundle list is largely composed of three-word strings, which account for 85% or 898 of the 1055 target bundles. They are followed by 147 four-word lexical bundles, which equal 14% of the total. There are only 10 different five-word lexical bundles in the corpus, representing 0.9% of all bundles. Tables 01, 02 and 03

https://doi.org/10.15405/epsbs.2019.09.46
Corresponding Author: Ang Leng Hong
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

display the most frequent three-word, four-word and five-word lexical bundles identified in IBM corpus in the descending order of normalised frequency (per million words=pmw).

Table 01. Top 20 three-word lexical bundles in order of normalised frequency

Rank	Frequency (pmw)	Three-word lexical bundle
1	452	more likely to
2	429	in order to
3	413	as well as
4	397	in terms of
5	370	the number of
6	366	the relationship between
7	344	the level of
8	319	the impact of
9	318	are more likely
10	296	the effect of
11	264	the effects of
12	250	the importance of
13	248	likely to be
14	222	the host country
15	220	in this study
16	216	as a result
17	212	the results of
18	209	based on the
19	204	the role of
20	197	are likely to

Table 02. Top 20 four-word lexical bundles in order of normalised frequency

Rank	Frequency (pmw)	Four-word lexical bundle
1	306	are more likely to
2	189	the extent to which
3	161	on the other hand
4	130	in the context of
5	120	in the host country
6	120	in the case of
7	104	on the basis of
8	88	the results of the
9	87	more likely to be
10	81	at the same time
11	77	as well as the
12	74	is positively related to
13	71	in terms of the
14	67	per cent of the
15	63	in the form of
16	62	is likely to be
17	60	it is important to
18	60	as a result of
19	58	to the extent that
20	56	more likely to have

https://doi.org/10.15405/epsbs.2019.09.46
Corresponding Author: Ang Leng Hong
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

Table 03. Top 20 five-word lexical bundles in order of normalised frequency

Rank	Frequency (pmw)	Five-word lexical bundle
1	55	are more likely to be
2	48	are more likely to have
3	42	firms are more likely to
4	42	the extent to which the
5	28	is positively related to the
6	28	the findings of this study
7	28	on the basis of the
8	24	the results of this study
9	21	in the context of the
10	20	they are more likely to

The results show that three-word lexical bundles, i.e., *more likely to, in order to, as well as, in terms of* and *the number of* are the most frequent lexical bundles in IBM corpus. Also, the most frequent three-, four- and five-word lexical bundles are *more likely to*, *are more likely to*, and *are more likely to be*, respectively. There is a possibility that the shorter lexical bundle (*more likely to*) is essentially part of the longer bundles (*are more likely to* and *are more likely to be*). With regard to this possibility, more detailed context analysis needs to be carried out to determine if the three-word lexical bundle is fully a fragment of the longer four- and five-word lexical bundles. Besides, it is apparent that the frequency and the length of lexical bundles are inversely related. This observation is in line with the characteristics of lexical bundles, that the longer the lexical bundle, the lower is its frequency (Biber et al., 1999; Hyland, 2008; Simpson-Vlach, & Ellis, 2010; Salazar, 2014).

7. Conclusion

The present study has employed corpus-based methods to identify three- to five- word lexical bundles commonly used in IBM research articles. Future research is needed to determine if the shorter lexical bundles are truly part of the longer lexical bundles and more detailed methodological criteria could be proposed in identifying lexical bundles of varying lengths. The present study suggests that EAP language instructors could take into account the lists of frequently used lexical bundles in IBM research articles to develop the relevant teaching materials for their courses. This study also contributes towards raising the awareness of the phraseological nature and tendency in academic writing, particularly in discipline-specific research articles. More phraseology research needs to be conducted to derive lists of phraseological sequences specific to different disciplines for EAP teaching and learning purposes.

Acknowledgments

This work was supported by Universiti Sains Malaysia Short Term Grant (304/PHUMANITI/6315044).

https://doi.org/10.15405/epsbs.2019.09.46
Corresponding Author: Ang Leng Hong
Selection and peer-review under responsibility of the Organizing Committee of the conference
eISSN: 2357-1330

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.

Anthony, L. (2015). *AntCont* (Version 3.4.1w) [computer software]. Available from http://www.laurenceanthony.net/software/antconc/releases/

Barlow, M. (2004). *Collocate (Version 1.0).* [computer software]. Available from http://www.athel.com/colloc.html

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Lexical expressions in speech and writing. In D. Biber & G. Leech (Eds.), *Longman grammar of spoken and written English* (pp. 988 -1036). Harlow, Essex: Longman.

Frankenberg-Garcia, A. (2018). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes, 35,* 93-104.

Green, C., & Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects. *Journal of English for Academic Purposes, 35,* 105-115.

Hsu, W. (2014). The most frequent opaque formulaic sequences in English-medium college textbooks. *System, 47,* 146-161.

Hyland, K. (2002). Specificity revisited: how far should we go now? *English for Specific Purposes, 21,* 385-395.

Hyland, K. (2006). *English for academic purposes: An advanced resource book*. New York: Routledge.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62.

Lei, L., & Liu, D. (2018). The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics, 23*(2), 216-243.

Liu, D. (2017). *Idioms: Description, comprehension, acquisition, and pedagogy*. UK: Routledge.

Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System, 42,* 220-231.

Salazar, D. (2014). *Lexical Bundles in native and non-native scientific writing*. Amsterdam: John Benjamins Publishing Company.

Shin, Y. K., Cortes, V., & Yoo, I. W. H. (2018). Using lexical bundles as a tool to analyze definite article use in L2 academic writing: An exploratory study. *Journal of Second Language Writing, 39,* 29-41.

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List (AFL). *Applied Linguistics, 31,* 487-512.

Wright, H. R. (2019). Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. *English for Specific Purposes, 54,* 1-14.