Journal of
Creative Practices in Language Learning and Teaching (CPLT)
Volume 5, Number 1, 2017

Multi-word Sequences in Learner Corpora: A Corpus Analysis of Lexical Bundles

ANG Leng Hong
lenghong@usm.my
School of Humanities
Universiti Sains Malaysia, MALAYSIA

HE Mengyu
mengyuhe1989@gmail.com
School of Humanities
Universiti Sains Malaysia, MALAYSIA

Received: 29 Dec 2016. Accepted: 13 Mar 2017 / Published online: 11 May 2017

ABSTRACT

There have been longstanding attempts to establish frequency profiles of words which are specific to academic register in order to facilitate learners in composing fluent academic writing. A more noteworthy effort was by Coxhead who proposed the Academic Word List (Coxhead, 2000). Recent developments in the field have increasingly regarded multi-word sequences such as lexical phrases, lexical bundles, formulas and clusters as crucially important and functionally significant in the academic contexts (Simpson-Vlach & Ellis, 2010). The present study adopts a corpus-based approach to identify a type of multi-word sequence, i.e., lexical bundles in student academic writing. Lexical bundles retrieved from a corpus of Asian college student essays and a corpus of British university-level student writing are identified, analysed and compared using corpus-linguistic techniques. The results of the analysis show that certain lexical bundles share the same keywords. This keyword sharing characteristic suggests that lexical bundles are internally analysable although they are initially retrieved as continuous strings of words. Besides, there is no significant difference in the functional use of lexical bundles between the Asian learners and British native students. However, both Asian learners and British university students are found to prefer different types of lexical bundles. Simpson-Vlach and Ellis's (2010) functional classification taxonomy (e.g., referential expressions, stance expressions, discourse organising functions) is used to categorise and analyse the items functionally. Finally this paper discusses the pedagogical implications drawn from the analysis.

Keywords: Corpus linguistics. Corpus-based. Multi-word sequence. Lexical bundles. Functional analysis

✉ He Mengyu
School of Humanities
Universiti Sains Malaysia, Penang MALAYSIA
E-mail: mengyuhe1989@gmail.com

INTRODUCTION

Multi-word sequences are important in language use and learning (Biber & Conrad, 1999; Wray, 2002; Biber, Conrad & Cortes, 2004; Schmitt, 2004; Simpson-Vlach & Ellis, 2010). Learning to use appropriate multi-word sequences contributes to a learner's communicative competence as research has shown that a dearth of knowledge of appropriate multi-word sequences leads to the increased and sustained mental processing burden which, in turn, could be a barrier to communication (Wray, 2000; Millar, 2011). The use of multi-word sequences has also been shown to be a vital measure of learner development (e.g., Pawley & Syder, 1983; Wray, 2002). Scholars have said that learners should master the use of appropriate multi-word sequences in language (Biber & Gray, 2013; Cortes, 2006). There are a variety of fixed and semi-fixed multi-word sequences which have often been referred to as formulaic sequences, collocations, idioms, formulas, prefabricated patterns, chunks, clusters, lexical bundles, recurrent sequences and n-grams (Nattinger & DeCarrico, 1992; Stubbs, 1995; Cowie, 1998; Manning & Schütze, 1999; Howarth, 1998; Biber et al. 1999; Wray, 2002; Schmitt, 2004). Hyland (2008) regards these multi-word sequences as "extended collocations which appear more frequently than expected by chance, helping to shape meanings and contributing to our sense of coherence in a text" (p.41). In academic settings, an important criterion that warrants effective and successful academic writing is the fluent control of multi-word sequences. It has been found that a significant proportion of academic discourse is made up of multi-word sequences, i.e., the recurrent lexical bundles (Biber et al., 1999). In recent years, there has been an increasing interest in establishing lexical and multi-word profiles of academic genres, for instance, the Academic Word List (Coxhead, 2000) and more recently the Academic Formulas List (Simpson-Vlach & Ellis, 2010) and the Academic Collocation List (Ackermann & Chen, 2013). This is due to the fact that academic study requires unique demands on language learners as patterns and constructions of academic register are different from those of other registers such as the conversational register which most learners are more familiar with. Thus, it concerns researchers if language learners master the use of multi-word sequences in order to succeed in academic writing.

PURPOSE OF THE STUDY

This paper presents a corpus-based study of a particular type of multi-word sequence, i.e., lexical bundles in student academic writing. Lexical bundles are "sequences of three or more words", self-contained within a clause with no structural completeness required as well as not idiomatic in nature (Biber et al., 1999, p. 990). The purpose of this study is to identify, analyse and compare lexical bundles extracted from two corpora of student academic writing, ***ICNALE*** and ***BAWE***. ***ICNALE*** is a compilation of Asian college student essays while ***BAWE*** is a collection of British university-level student writing. This study is motivated by the findings put forward by Cortes (2004) and Salazar (2014) in which non-native learners are found to overuse certain lexical bundles such as discourse organising lexical bundles and at the same time, they underuse other lexical bundles such as referential lexical bundles which are commonly found in native writing. As far as the researchers are concerned, no attempts have been made to compare academic essays produced by groups of Asian learners with those of British native students. This study therefore aims at examining the similarities or differences in the use of lexical bundles by these two groups of students.

The lexical bundles identified are categorised and analysed based on Simpson-Vlach and Ellis's (2010) functional classification (e.g., referential expressions, stance expressions, discourse organising functions). It should be noted that Simpson-Vlach and Ellis (2010) adapted the functional classification from Biber et al. (2004). There is another functional classification proposed by Hyland (2008). However, the present study employed Simpson-Vlach and Ellis's (2010) functional classification taxonomy to classify lexical bundles functionally as this taxonomy is deemed more suitable for the classification of lexical bundles produced by students. The study focuses on comparing the functional use of the lexical bundles in two learner corpora as lexical bundles are essentially functionally operative in the texts (Biber et al., 2004). Functional analysis of lexical bundles is essential to their value from the pedagogical perspective. Lexical bundles can be used to introduce topics, compare and contrast ideas and draw conclusions. Lists of useful lexical bundles therefore could be incorporated into the syllabus of English for Academic Purposes (EAP) courses.

Table 1 presents the functional taxonomy developed by Simpson-Vlach and Ellis (2010). As shown in Table 1, there are three main functional categories of lexical bundles. According to Biber et al. (2004), referential expressions generally identify entity or attribute of an important entity; Stance bundles are useful in conveying epistemic meaning and writer's attitude towards a particular proposition; Discourse organising bundles are mainly functioned as topic introduction and elaboration. Discourse organising bundles are also useful to signal or refer to prior or upcoming discourse (Simpson-Vlach & Ellis 2010).

Table 1
Taxonomy for functional classification of lexical bundles proposed by Simpson-Vlach and Ellis (2010, p. 498-502), adapted from Biber et al. (2004)

Simpson-Vlach and Ellis (2010)		***Example***
Referential expression	• Specification of attributes	
	➢ Intangible framing attributes	*in relation to, in the context of*
	➢ Tangible framing attributes	*the sum of, the size of the*
	➢ Quantity specification	*a set of, a large number of*
	• Identification and focus	*different types of, that is the*
	• Contrast and comparison	*the difference in, the same as*
	• Deictics and locatives	*the real world, at this stage*
	• Vagueness markers	*and so forth, and so on*
Stance expression	• Hedges	*likely to be, it appears that*
	• Epistemic stance	*be regarded as, can be considered*
	• Obligation and directive	*it should be noted, need to be*

Discourse organising function	• Expressions of ability & possibility	*can be used to, are able to*
	• Evaluation	*it is important, is consistent with*
	• Intention/Volition	*to do so, do not intend to*
	• Metadiscourse & textual reference	*In the next section, in this paper*
	• Topic introduction and focus	*for example in the, first of all*
	• Topic elaboration	
	➢ Non-causal	*factors such as, are as follows*
	➢ Cause and effect	*as a result of, due to the*
	• Discourse markers	*at the same time, in other words*

RESEARCH QUESTIONS

This study examined the lexical bundles employed by Asian college learners and British university students in their academic essay writing to shed light on the following questions:

1. Which lexical bundles are found in:
    a) Asian college students' academic essays?
    b) British university students' academic essays?
2. To what extent do the lexical bundles employed by Asian college students and British university students differ functionally?

CORPORA AND METHODS

The data analysed in the study were drawn from "The International Corpus Network of Asian Learners of English" (***ICNALE***) and "The British Academic Written English Corpus" (***BAWE***). The ***ICNALE*** is a collection of academic essays written by college students in 10 Asian countries and areas, namely China Mainland, Taiwan, Hong Kong, Singapore, Thailand, Pakistan, India, The Philippines, Japan and Korea. The proficiencies of all these Asian college students were tested and classified into 4 levels, namely waystage, lower, upper and vantage. The present study explored the ***ICNALE*** as a whole by looking at the lexical bundles produced by Asian learners in general in order to compare the lexical bundles with those found in ***BAWE***. It did not intend to look at the use of lexical bundles by learners of various proficiency levels. The ***ICNALE*** contains 5200 texts and comprises 1.2 million words. The text lengths in the ***ICNALE*** range from 200 to 300 words. The contributors of the ***BAWE*** are proficient British university-level students from four broad disciplinary areas (Arts and Humanities, Social Sciences, Life Sciences and Physical Sciences). Only a sub-corpus of the ***BAWE*** was used in the present study

in order to achieve comparability with regard to the sizes of the corpora used in this study. Comparisons work much better when the corpora being compared are of similar size (Rayson, 2003). The sub-corpus of the ***BAWE*** totalling 1.2 million words is composed of 469 essays which are produced by proficient British university students in Britain. The text lengths in the ***BAWE*** range from 500 to 5000 words.

The corpora were explored using *AntConc* software version 3.4.1w, a corpus processing tool developed by Anthony (2014) which facilitates the analysis of lexical bundles or n-Grams. To answer Research Question 1, *n-Grams tool* in *AntConc* was used to identify and extract 4- and 5-grams (4-word and 5-word lexical bundles) occurring at least 30 times per million words in each corpus. According to the literature, the minimum cut-off points range from 10 to 40 times per million words (Biber et al., 1999; Cortes, 2004; Simpson-Vlach & Ellis, 2010). In the present study, it was decided to have the minimum cut-off point set at 30 times per million words. This is to ensure that only lexical bundles that are used very frequently by the learners were extracted by *AntConc* software. Manual filtration was carried out by retaining lexical bundles which serve discourse-pragmatic functions in the texts. Subject-specific bundles were omitted from the candidate lists. The lexical bundles extracted from both corpora that qualify under the criteria were shown in Appendix.

To answer Research Question 2, the identified lexical bundles were analysed and classified based on Simpson-Vlach and Ellis's (2010) functional classification (e.g., referential expressions, stance expressions, discourse organising functions). Concordance analysis was required in which the lexical bundles were put back into their original textual contexts in order to determine the discourse functions they serve accurately. Some of the lexical bundles are multifunctional. However, only one discourse function deemed most probable was assigned to each lexical bundle.

FINDINGS AND DISCUSSION

LEXICAL BUNDLES FOUND IN ICNALE AND BAWE

The results in the corpus analysis are normalised to per million words for the purpose of comparability. Table 2 shows the top 10 lexical bundles in each corpus that qualify under the criteria found in both the ***ICNALE*** and the ***BAWE***. Please refer to Appendix for full lists of lexical bundles.

Table 2
Top 10 lexical bundles in the ***ICNALE*** and the ***BAWE***

No	Lexical bundles in ICNALE	No. of lexical bundles per million words	No	Lexical bundles in BAWE	No. of lexical bundles per million words

1	It is important for	773	*1*	*As a result of*	*153*
2	That it is important	325	*2*	*On the other hand*	*136*
3	I agree with the	231	*3*	*In the case of*	*95*
4	On the other hand	212	*4*	*The end of the*	*94*
5	Think that it is	193	*5*	*It is important to*	*87*
6	At the same time	188	*6*	*The fact that the*	*85*
7	I think that it	174	*7*	*As well as the*	*83*
8	I agree with this	169	*8*	*In the form of*	*71*
9	Is one of the	150	*9*	*It is clear that*	*70*
10	I think that it is	148	*10*	*In terms of the*	*63*

There are a total of 92 types of relevant lexical bundles found in Asian learners' essays, while 43 are discovered in British university students' writing (refer to Appendix). A manual inspection of the lists of lexical bundles derived from both corpora reveals that some four-word lexical bundles could be subsumed into the relevant longer five-word bundles while some could not because the shorter four-word bundles co-occur with other preceding or succeeding collocates apart from the ones in the longer five-word bundles. Besides, it is worth noting that some lexical bundles show certain structural and semantic affinities. Their structural and semantic relationships are detected when these lexical bundles are found to have the same keywords. In this context, the *keyword* refers to the word that is central to the whole bundle. The use of keyword in examining lexical bundles was initially proposed by Salazar (2014) who examined lexical bundles in native and non-native scientific writing. In the current study, a closer inspection of concordance lines reveals that most of the frequent four- and five-word bundles have the same keywords, which are in the forms of verb and adjective. For instance, the lexical bundles, *I think it is*, *think it is a*, *think that it is*, *I think that it is*, *I think that the*, *I think it is very*, *I think this is* share the same keyword *think*. Besides verb, frequent lexical bundles found in both corpora also share the same adjective, such as *important*, as in the following lexical bundles: *it is important for*, *is important for us*, *that it is important*, *is very important for*, *it is important to*, *it the most important*, *most important thing for*, *most important thing is.* In view of the keyword sharing characteristic of lexical bundles, this finding shows that lexical bundles are in fact internally analysable, although they were initially identified and retrieved as continuous multi-word sequences. Many previous studies on lexical bundles derive only lists of bundles or other types of multi-word sequences such as collocations in frequency order (e.g., Shin & Nation, 2008; Durrant, 2009; Simpson-Vlach & Ellis, 2010; Hsu, 2014). The finding on keyword sharing

characteristic has provided new perspective for approaching and organising the frequent lexical bundles, particularly in the academic context.

THE FUNCTIONAL USE OF LEXICAL BUNDLES BY ASIAN COLLEGE STUDENTS AND BRITISH UNIVERSITY STUDENTS

The lexical bundles were further analysed for the purpose of functional classification. Tables 3 and 4 below show the total number and percentages of lexical bundles as well as examples of lexical bundles extracted from the ***ICNALE*** and ***BAWE,*** respectively. The lexical bundles were classified according to Simpson-Vlach and Ellis's (2010) functional classification taxonomy.

Table 3
Lexical bundles classified according to functions in the ***ICNALE***

No.	Discourse functions	Type of Lexical bundles	Frequency and percentages of lexical bundles
1	Referential expressions		
	a) Specification of attributes	*my point of view, there are lots of*	5 (5.4%)
	b) Identification and focus	*that it is very, if there is a*	7 (8%)
2	Stance expressions		
	a) Epistemic stance	*I agree with the, I think this is*	26 (28%)
	b) Expressions of ability and possibility	*will be able to, to be able to*	3 (3.3%)
	c) Evaluation	*it is very important for, it is hard to*	40 (43.5%)
	d) Intention/volition	*I would like to, if we want to*	7 (7.6%)
3	Discourse organising functions		
	a) Discourse markers	*at the same time, as well as the*	2(2.1%)
	b) Topic elaboration	*because it is, that is why I*	2 (2.1%)
	Total		**92 (100%)**

Table 4
Lexical bundles classified according to functions in the ***BAWE***

No.	Discourse functions	Instances of Lexical bundles	No. of instances and percentages of combined lexical bundles
1	Referential expressions		
	a) Specification of attributes	*in the case of, in terms of the*	21 (49%)
	b) Identification and focus	*that there is a, in this case the*	4 (9.3%)
	c) Contrast and comparison	*the relationship between the*	1 (2.3%)
2	Stance expressions		
	a) Epistemic stance	*it can be argued that, we can see that*	4 (9.3%)
	b) Expressions of ability and possibility	*it is possible to, can be seen in*	1 (2.3%)
	c) Evaluation	*it is important to, it is clear that*	5 (11.6%)
3	Discourse organising functions		
	a) Topic elaboration	*as a result of the, as a consequence of*	2 (4.6%)
	b) Discourse markers	*as well as the, at the same time*	5 (11.6%)
	Total		**43 (100%)**

In order to find out the significant differences of the use of lexical bundles between the two groups of learners (***ICNALE*** and ***BAWE***), Mann-Whitney U test was performed to calculate the U-value. The U-value is 28.5. The critical value of U at $p < .05$ is 17. Therefore, the result is not significant at $p < .05$. This implies that both Asian learners and British university students do not differ significantly in using the lexical bundles functionally. There is no concrete evidence in showing the functionally overuse or underuse of lexical bundles by both groups of learners. This result is contradictory to the findings reported by Cortes (2004) and Salazar (2014) in which non-native learners are found to overuse certain lexical bundles such as discourse organising lexical bundles and at the same time, they underuse other lexical bundles such as referential lexical bundles which are commonly found in native writing.

However, the descriptive statistic seems to suggest that Asian college students (in the ***ICNALE***) and British university students (in the ***BAWE***) prefer different types of lexical bundles. As shown in Tables 3 and 4, of all types of lexical bundles in the ***ICNALE***, 82.4% of the lexical bundles serve as stance expressions while 60.6% of the types of lexical bundles in the ***BAWE*** are found to be referential expressions. Lexical bundles serving as discourse organisers appear to be the smallest group in both the ***ICNALE*** and the ***BAWE***. On the whole, it seems evident that the Asian college learners prefer stance expressions in their academic writing while British university students tend to employ more referential expressions in their essay writing.

With regard to stance expressions, epistemic expressions are usually used to make knowledge claims and express beliefs as well as opinions (Biber et al., 2004; Simpson-Vlach & Ellis, 2010), while evaluation bundles express evaluative meanings (Hunston, 2011). In ***ICNALE***, it has been found that epistemic stance and evaluation are very frequently used by Asian learners. On the other hand, these stance expressions are less preferred by British native students in writing for academic purposes.

The corpus analysis shows that referential expressions are predominant in ***BAWE***. The sub-category, specification of attributes bundles are frequently employed by British native students in writing their academic essays. According to Biber et al. (2004), these bundles typically identify specific attributes of the succeeding head noun. It is worth mentioning that in ***ICNALE***, referential bundles are rarely used by Asian learners. This finding is partly in line with the results of Chen and Baker (2010), that non-native students use proportionally fewer referential expressions than the professional writers of published academic writing. It is also crucial to note that referential expressions are commonly used in academic writing as academic writing requires the identification of important entities and attributes (Biber et al., 2004) for corroboration and validation purposes.

CONCLUSION AND PEDAGOGICAL IMPLICATIONS

This study has produced some interesting findings. First, lexical bundles are found to show certain structural and semantic affinities. Their structural and semantic relationships are detected when these lexical bundles share the same keywords. The keyword sharing characteristics of lexical bundles indicate that lexical bundles are internally analysable, although they are initially retrieved as continuous strings of words. Second, Asian college learners and British university students do not differ significantly in using the lexical bundles functionally. There is no strong evidence which validates the claim that non-native (Asian learners) overuse or underuse certain types of lexical bundles, as reported in the previous studies. Third, the descriptive statistical analysis suggests that both Asian learners and British university students prefer different types of lexical bundles in writing academic essays. It should be noted that the preference of certain types of lexical bundles does not imply the overuse of certain types of lexical bundles. The notion of "use more" is different from "overuse" technically and statistically. Fourth, the findings also indicate that lexical bundles are prevalent in both native and non-native student writing. It can be safely said that lexical bundles are pervasive in language, particularly in the academic settings. To reiterate, scholars in the field have pointed out that an important criterion that warrants

effective and successful academic writing is the fluent control of multi-word sequences. Thus, more emphasis should be given to the teaching and learning of multi-word sequences such as lexical bundles in the language courses.

The results of analysis also show that Asian college learners do not favour referential expressions in their writing. As mentioned earlier, referential bundles are important in the academic settings. Lexical bundles which serve referential purposes could therefore be given more emphasis in language classrooms. Besides, the ***ICNALE*** does not include Malaysian learner writing. Malaysian learners, in particular, may be taught with lexical bundles in the classroom to improve the standard of their English due to the persuasive nature and the importance of lexical bundles in academic writing. Nevertheless, learning lists of lexical bundles is not sufficient; learners should be guided on how to use lexical bundles in the academic context effectively. That is one of the reasons the current study also analysed the discourse functions of lexical bundles in both learner corpora. With a better grasp of the discourse functions of lexical bundles, non-native learners, particularly the Malaysian learners will be able to know how to use lexical bundles in the academic context more efficiently.

In terms of how to implement the teaching of lexical bundles in classroom, findings from this study suggest that adopting lexical bundles marked by referential functions is valued in both sets of native and non-native writings. Thus, the selection of referential bundles could benefit the language teaching. Previous studies also suggest that the teaching of multi-word sequences could focus on highly frequent bundles (e.g., Cortes, 2006; Hyland, 2008; Jones & Haywood, 2004). As lexical bundles are varied across different disciplinary contexts, "disciplinarity and specialisation" (Eriksson, 2012) should also be considered when deciding which bundles to focus on in a teaching situation. Future research may delve into types of exercises on lexical bundles aligned with different learner levels. Academic lexical bundles such as the recent *Academic Formulas List* proposed by Simpson-Vlach and Ellis (2010) can be used by instructors in the courses of EAP as teaching materials. In sum, language learners, particularly in higher institutions of learning should be exposed to more multi-word sequences such as lexical bundles as this will facilitate them in composing fluent academic writing.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers and CPLT editors for their helpful comments. The co-author is a PhD candidate at the School of Humanities, USM. Her research is supported by USM Fellowship.

The data in this study come from:

1) The ***ICNALE***: The International Corpus Network of Asian Learners of English

The ***ICNALE*** is a collection of 1.2M words of controlled essays written by English learners in 10 countries and areas in Asia.

Project Leader: Dr. Shin'ichiro Ishikawa, Kobe University, Japan.

2) The ***BAWE***: the British Academic Written English corpus

The ***BAWE*** was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics [previously called CELTE], Warwick), Paul Thompson (Department of Applied Linguistics, Reading) and Paul Wickens (Westminster Institute of Education, Oxford Brookes).

REFERENCES

Ackermann, K & Chen, Yu-Hua. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*, 235-247.

Anthony, L. (2014). *AntConc version 3.4.1w*. Japan: Waseda University.

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

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose, in Hasselgard & Oksefjell. *Out of corpora: Studies in honour of Stig Johansson*. Atlanta: Rodopi Press.

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

Biber, D., & Gray, B. (2013). *Discourse characteristics of writing and speaking task types on the TOEFL iBT test: A lexico-grammatical analysis (TOEFL iBT Research Report No. 19)*. Princeton, NJ: Educational Testing Service.

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

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397–423.

Cortes, V. (2006). Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, *17*, 391-406.

Cowie, A. P. (1998). *Phraseology: Theory, analysis, and applications*. Oxford: Oxford University Press.

Coxhead, A. (2000). A new Academic Word List. *TESOL Quarterly*, *34*, 213-38.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for Academic Purposes. *English for Specific Purposes*, *28*(3), 157-169.

Eriksson, A. (2012). Pedagogical perspectives on bundles: Teaching bundles to doctoral students of biochemistry. In J. Thomas & A. Boulton (Eds). *Input, Process and Product: Developments in Teaching and Language Corpora* (pp. 195-211). Brno: Masaryk University Press.

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

Howarth, P. (1998). The phraseology of learners' academic writing. In A.P. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 161-186). Oxford: Oxford University Press.

Hunston, S. (2011). *Corpus approaches to evaluation: Phraseology and evaluative language*. New York: Routledge.

Hyland, K. (2008). As can be seen: lexical bundles and disciplinary variation. *English for Specific Purposes*, *27* (1), 4-21.

Jones, M., & S. Haywood. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 269-292). Amsterdam/ Philadelphia: John Benjamins.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Millar, N. (2011). The processing of malformed formulaic language. *Applied Linguistics*, *32*(2), 129-148.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

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). New York: Longman.

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

Shin, D., & Nation, I.S.P. (2008). Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, *62*(4), 339-348.

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

Stubbs, M. (1995). Collocations and semantic profiles. On the cause of the trouble with quantitative studies. *Functions of Language*, *2*(1), 23-55.

Schmitt, N. (2004). *Formulaic sequences: Acquisition, processing and use*. Amsterdam: John Benjamins.

Rayson, P. (2003). *Matrix: A Statistical Method and Software Tool for Linguistic Analysis through Corpus Comparison*. PhD dissertation. Lancaster University.

Wray, A. (2000). Formulaic sequences in second language teaching: principle and practice. *Applied Linguistics*, *21*(4), 463-489.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

About the Authors

Dr. Ang Leng Hong is currently teaching at the English Language Studies section, School of Humanities, Universiti Sains Malaysia. Her research interests include corpus linguistics, vocabulary and phraseology.

He Mengyu is currently a PhD candidate at the English Language Studies section, School of Humanities, Universiti Sains Malaysia. Her research interests include meta-discourse, lexical bundles and corpus linguistics.

APPENDIX

No.	Lexical bundles in the *ICNALE*	Raw Freq.	Lexical bundles in the *BAWE*	Raw Freq.
1	It is important for	928	As a result of	184
2	That it is important	390	On the other hand	163
3	I agree with the	277	In the case of	114
4	On the other hand	254	The end of the	113
5	Think that it is	232	It is important to	104
6	At the same time	225	The fact that the	102
7	I think that it	209	As well as the	99
8	I agree with this	203	In the form of	85
9	Is one of the	180	It is clear that	84
10	I think that it is	178	In terms of the	75
11	Is not good for	172	It is possible to	74
12	Agree with the statement	168	A result of the	66
13	Is very important for	141	That there is a	64
14	It is important to	141	At the same time	62
15	I agree with the statement	135	At the end of	60
16	The most important thing	131	It is difficult to	60
17	Is a good way	116	As a result of the	59
18	There are a lot	116	Can be seen in	57
19	A good way to	115	One of the most	56
20	There are a lot of	112	To the fact that	53
21	Agree that it is	111	The nature of the	52
22	Will be able to	107	The extent to which	49
23	It is a good	106	Is one of the	48
24	If you want to	104	The role of the	48
25	It is very important	102	The war of the	48
26	Agree that it is important	100	In this case the	47
27	Think that it is important	95	The rest of the	47
28	It is not good	94	The development of the	46
29	Is the most important	93	In the absence of	44
30	I agree that it	91	The relationship between the	44
31	I agree with this statement	90	It can be argued	42
32	I agree that it is	84	At the time of	41
33	It is good for	83	By the fact that	41
34	So I think it	81	On the basis of	41
35	Is a good way to	77	The majority of the	40
36	Is not a good	75	As a consequence of	39

37	It is not a	73	The importance of the	39
38	Do not want to	72	We can see that	39
39	I think that the	71	it could be argued	38
40	In my opinion it	70	At the end of the	37
41	To be able to	68	Can be argued that	36
42	It is hard to	62	It can be argued that	36
43	It is not only	61	as a means of	36
44	I do not agree	59		
45	It is true that	59		
46	It is difficult to	58		
47	I would like to	56		
48	The best way to	56		
49	There are so many	56		
50	One of the most	55		
51	My point of view	53		
52	It is not good for	52		
53	It is necessary for	52		
54	As well as the	51		
55	It is said that	49		
56	I do not think	48		
57	Most important thing is	48		
58	I disagree with the	47		
59	I strongly agree that	47		
60	Is very bad for	47		
61	Think it is a	46		
62	Is the best way	46		
63	I think it is not	46		
64	The most important thing is	45		
65	Is a good idea	45		
66	They want to do	45		
67	If we want to	44		
68	Not be able to	44		
69	Think it is good	43		
70	Is it important for	43		
71	It is not easy	43		
72	There are lots of	43		
73	I think it is very	42		
74	I think this is	42		
75	Because it is a	41		

76	Is very important to	41
77	So it is important	40
78	I believe that it	40
79	It is necessary to	40
80	That it is very	40
81	Is important for us	39
82	Is the most important thing	39
83	What they want to	39
84	Is more important than	38
85	Most important thing for	37
86	If there is a	37
87	If you do not	37
88	I think it is	36
89	Do not think that	36
90	Is not easy to	36
91	That is why I	36
92	The number of people	36