3L: The Southeast Asian Journal of English Language Studies – Vol 24(2): 82 – 94
http://doi.org/10.17576/3L-2018-2402-07

Specificity in English for Academic Purposes (EAP): A Corpus Analysis of Lexical Bundles in Academic Writing

ANG LENG HONG
School of Humanities
Universiti Sains Malaysia
lenghong@usm.my

TAN KIM HUA
Faculty of Social Sciences & Humanities
Universiti Kebangsaan Malaysia

ABSTRACT

The issue of specificity in English for Academic Purposes (EAP) settings has always challenged linguists and instructors in the field to take a stance on how language should be perceived, that is whether language forms and features are transferable across different academic disciplines or are specific to particular disciplines. This study intends to take this debate a step further by employing a corpus-driven method in identifying a type of phraseological sequence, namely lexical bundles in a corpus of journal articles in the field of International Business Management (IBM). The lexical bundles were compared with those compiled by Simpson-Vlach and Ellis (2010) in their study of Academic Formulas List (AFL) to determine the specificity of the lexical bundles identified in this study. Following frequency-based approach, the corpus tool, Collocate 1.0 was used to extract three- to five-word sequences. These word sequences were manually filtered to exclude irrelevant and meaningless combinations. The qualified lexical bundles were compiled and compared with lexical bundles in AFL (Simpson-Vlach and Ellis 2010) using log-likelihood test. The findings show that three-word lexical bundles are the most common types of lexical bundles in IBM corpus. The comparison reveals that lexical bundles in IBM corpus are relatively specific as compared with lexical bundles in AFL. A discipline-specific approach to the teaching and learning of lexical bundles in EAP settings is therefore advocated to enhance EAP syllabuses and instruction.

Keywords: EAP; phraseological sequences; lexical bundles; frequency-based; discipline-specific

INTRODUCTION

Studies on phraseology in various genres and disciplines have been flourishing in recent years with the advancement of computer-mediated research methodology. Phraseology has been studied under the rubrics of, for instance, chunks, phraseological sequences, formulaic language, lexical bundles, collocations, multi-word items, recurrent sequences, n-grams, lexical phrases, and so on. Previous studies on phraseology have shown that the knowledge of phraseology is essential in ensuring fluency and natural use of language (Pawley & Syder 1983, Sinclair 1991, Hill 2000, Hyland 2012, Ang et al. 2017). Also, the appropriate use of phraseological sequences is a determining factor in warranting pragmatic competence, given the prevalence of these recurring sequences in both spoken and written discourse (Paquot & Granger 2012). The prevalence of phraseological sequences in discourse indicates that meaning creation and understanding is essentially dependent upon stocks of the phraseological sequences in language users' lexicon. In academic discourse, the mastery of the relevant phraseological sequences is particularly important to learners so that they could have access to the relevant "academic community" (Coxhead 2008, p. 151). Nevertheless, the formal conventions of academic discourse that are markedly different from those of other genres such as the conversational one pose difficulties for learners in processing information and interacting within the academic community in which they are in. Attention has thus been

given to the learning of academic conventions in the English for Academic Purposes (henceforth EAP) courses.

LITERATURE REVIEW

TWO APPROACHES TO EAP

The literature review section includes the review of the approaches to EAP before looking at the debates revolving around the issue of specificity in EAP settings. The inclusion of specific phraseological sequences in EAP curriculum is a debatable issue as there are essentially two approaches to EAP, i.e., the common-core approach and the discipline-specific approach. The common-core approach to EAP focuses on phraseological sequences common to all disciplines (Simpson-Vlach & Ellis 2010, Schutz 2013). The discipline-specific approach concerns the degree of specificity of the phraseological sequences in different disciplines (Cortes 2004, Hyland 2006, Durrant 2014). The advocates of this discipline-specific approach hold on to the claim that there are significant amount of formalities in academic texts, which are characterised by the use of subject-specific phraseological expressions. The distinction between the practices of these two approaches is also widely known as English for General Academic Purposes (EGAP) and English for Specific Academic Purposes (ESAP) (Hyland 2006). Technically, EGAP is "concerned with the general academic language and study skills" that are common across different academic disciplines whereas ESAP "is concerned with the language features of particular academic disciplines or subjects" (Jordan 1989, p. 151).

The issue of specificity has been debated among the scholars who hold different views on the approaches to EAP. Some EAP writers, such as Hutchison and Waters (1987), Spack (1988) and Zamel (1993), strongly argue against discipline-specific teaching based on several reasons. First, EAP teachers are not trained to handle subject-specific forms of language, and they do not possess the expertise to teach specialist contents. Spack (1988) proposes that these discipline-specific conventions should be taught by subject teachers themselves as they know these specialist contents best. Second, in EAP classrooms, the main focus is generic and literacy skills, including making paraphrases and summaries as well as giving oral presentations during tutorial classes and seminars. These activities are said to differ little across the disciplines (Jordan 1997). Last, there is the idea which underlies all the others: that there are forms of language that transcend disciplinary boundaries and EAP teachers should adopt a common-core approach to teach "general principles of inquiry and rhetoric" (Spack 1988, p. 29) in language classrooms. Also, Hutchison and Waters (1987) claim that there are insufficient variations in various language forms and functions of different academic subjects to justify a discipline-specific approach. In this sense, a milder stance on the common-core approach to EAP teaching is taken by some writers, who concede the fact that different academic disciplines may show variations. Nevertheless, these writers maintain that "besides these discipline-specific features, there remains a teachable common core" (Coxhead 2000, 2008, Granger & Paquot 2009, p. 101). They propose that the discipline-specific features of EAP can be highlighted by EAP instructors when needed. With regard to the teaching of phraseological sequences in EAP classrooms, Simpson Vlach and Ellis (2010) suggest that a general approach to EAP is sufficient to elicit lists of common core academic clusters that transcend disciplinary boundaries. In their study of academic formulas, they were able to derive frequent lexical bundles which are common to many academic disciplines and are of general academic use.

In response, there are several justifications made to defend the ESAP position. First, to counter the position that discipline-specific language should be taught by lecturers in the relevant disciplines, Hyland (2002, 2006) argues that subjects specialists usually do not emphasise the generic and language skills in lectures due to two main reasons. Firstly, subject specialists are not trained to teach language and they generally "lack both the expertise and desire to teach literacy skills" (Hyland 2002, p. 388). Secondly, it appears that many lecturers in various disciplines consider academic discourse conventions as "largely self-evident and universal" (ibid.). Subject lecturers often assess students' work without concerning much with how the language conventions and forms are used (Braine 1988, Lea & Street 1999, Hyland 2002, 2006). It is worth noting that the responsibility of teaching language conventions and skills lies ultimately with EAP teachers as they are trained to handle language classrooms. To cope with the diverse requirements and needs of learners from various academic disciplines, EAP instructors should be trained in a more professional way to teach specialised language used in different academic disciplines or domains.

Second, the claim that EAP courses mainly focus on generic skills such as summarising and paraphrasing as well as making presentations which are not much varied across the different disciplines deserves a second thought. It should be borne in mind that the main goal of setting up EAP courses is to prepare learners with specific language skills relevant to their respective disciplines (Hyland 2002). EAP teachers should primarily concentrate on the teaching of language forms that carry distinctive and "clear disciplinary values" (Hyland 2006, p. 12) which are frequent and important to the relevant discourse community. The teaching of the relevant phraseological expressions deserves to be prioritised in EAP classrooms as these phraseological expressions such as lexical bundles are the "basic building block of discourse" in academic writing (Biber et al. 2004, p. 371).

Lastly, it is disputable that there is a common core of language items. Hyland and Tse (2007, p. 238) doubt that there is "a single inventory [that] can represent the vocabulary of academic discourse and so be valuable to all students irrespective of their field of study". With the development of corpus-based studies in recent years, studies on vocabulary and phraseological sequences have been able to inform the necessary vocabulary and phrases teaching in EAP. These studies evidently show that there are significant variations between disciplines (Cortes 2004, Hyland & Tse 2007, Hyland 2008a, Durrant 2014). In addition, the variations between genres and registers have also been studied and proven to be a reality in the academic settings (Biber et al. 1999, Biber et al. 2004, Hyland 2008b). Also, any language forms may possibly have a number of different meanings and functions depending on the contexts in which the language is used. It is therefore sensible to claim that vocabulary behaves differently across disciplines and contexts (Hyland 2002, 2006). In a more assertive tone, Hyland and Tse (2007, p. 240) state that "all disciplines shape words for their own uses" and thus defend the discipline-specific approach to EAP.

The debate concerning which approaches should be established in EAP still continues as the rapid development of corpus linguistics continues to inform language teaching in EAP. The issue of specificity can impact the way EAP practitioners see the field and how they carry out their teaching. More studies need to be carried out to ascertain if the issue of specificity applies to the teaching of useful phrases in EAP classrooms. This study intends to take this debate a step further by comparing two lists of phraseological sequences which are compiled for the purposes of EGAP and ESAP, respectively.

PURPOSE OF THE STUDY

In order to see how language should be perceived and informed in the EAP settings, this study compares lists of phraseological sequences derived from two approaches (ESAP and EGAP). Specifically, this study attempts to identify a type of phraseological sequence, i.e. lexical bundles from a specialised corpus of journal articles in the field of International Business Management (henceforth IBM). The lexical bundles identified are compared with the lexical bundles in the Academic Formulas List (henceforth AFL) (Simpson-Vlach and Ellis 2010) to determine the specificity of the lexical bundles in this study. Following common-core approach, AFL (Simpson-Vlach & Ellis 2010) is a list of EGAP lexical bundles retrieved from a corpus of academic writing sampled across four academic disciplines: Humanities and Arts, Social Sciences, Natural Sciences /Medicine and Technology and Engineering while the lexical bundles identified in this study represent ESAP lexical bundles extracted from a specialised corpus which contains only research articles relevant to the field of IBM.

METHOD

The corpus and methods used to identify the discipline-specific lexical bundles are described in the following sub-sections.

THE CORPUS

The corpus for this study consists of academic research articles in the field of IBM. The journal articles were selected and compiled electronically. The selection of journals was based on the impact factor of the journals recognised by Thomson Reuters Web of Science. A total of two international journals were chosen. The rationale for selecting these journals is due to their specificity in publishing research articles pertaining to the field of IBM. The corpus consists of 1 million word tokens, and it includes 138 original research articles.

THE CORPUS TOOL

The corpus tool, *Collocate 1.0* (Barlow 2004) was used to extract lexical bundles automatically by setting the span options. This corpus tool recognises plain text files which end with *.txt* extension. *Collocate 1.0* extracts lists of n-grams (lexical bundles) using two statistical measures: frequency and Mutual Information.

STEPS IN IDENTIFYING LEXICAL BUNDLES

The first step of the analysis was to create a list of the most frequent lexical bundles of IBM. In accordance with Biber et al. (1999), lexical bundle is defined in this study as a frequently recurring sequence of words. As lexical bundles are a type of phraseological sequence, the terms *lexical bundles* and *phraseological sequences* are used interchangeably in this study. Following Biber et al. (1999), this study focuses on three- to five-word lexical bundles. The steps taken in identifying and determining the eligibility of phraseological sequences as lexical bundles are shown in Figure 1.

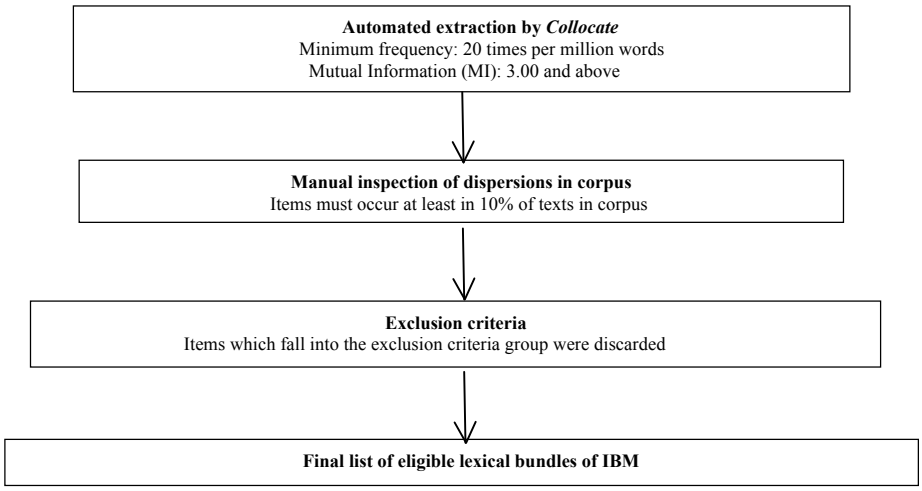

FIGURE 1. Steps in identifying lexical bundles

The lexical bundles were identified using the frequency-based approach. There was a minimum cut-off point for retrieving the lexical bundles (Biber et al. 1999). Another important statistic used to create the list of lexical bundles is the Mutual Information (MI) score. MI is a measure of the strength of association between words. A higher MI score means a stronger association and thus a more coherent relationship between words (Simpson-Vlach & Ellis 2010, Salazar 2014). This metric was applied in order to eliminate those word sequences that do not have meaning or function but occur often because of the high frequency of words that they contain. It was also used to avoid discounting useful but less frequent phrases that tend to end up at the bottom of frequency-based lists (Simpson-Vlach & Ellis 2010). Also, the dispersion criterion is necessary to avoid individual writers' idiosyncrasies (Hyland 2008b).

Collocate 1.0 extracted a total of 1714 three-word sequences, 270 four-word sequences and 25 five-word sequences. After the extraction by *Collocate 1.0*, the next step was to check the dispersions of phraseological sequences in corpus. A phraseological sequence has to occur in 10% of texts to avoid idiosyncrasies of particular writers (Hyland 2008b). It was discovered that not every phraseological sequence on the list was of phraseological relevance and therefore further sifting was necessary in order to produce a more refined list of lexical bundles.

Following Salazar (2014), some exclusion criteria were adapted in order to weed out irrelevant word combinations. The modified criteria and some instances of excluded bundles are shown in Table 1 below.

TABLE 1. Exclusion criteria for irrelevant word combinations

1)	Fragments of other bundles : *on the basis* (*On the basis of*), *in the case* (*in the case of*)
2)	Bundles consisting acronyms: *gdp per capita, OECD anti-bribery convention*
3)	Bundles composed exclusively of function words: *have also been*, *as it is*
4)	Bundles with random numbers : *at least one*, *for the first*
5)	Random section titles : *fig 1 b*, *table 2 in*
6)	Meaningless bundles: *it that is*, *studies e g*
7)	In-text citations : *Beck et al. , Gatignon Anderson 1988*

After excluding the irrelevant word combinations, the remaining lexical bundles were identified and arranged according to normalised frequency order (per million words). The most frequent lexical bundles in this study were compared with those of Simpson-Vlach and Ellis's (2010) study to determine the specificity of the lexical bundles in this study. A statistical measure, log-likelihood test was performed on the lexical bundles found in both

studies. The results of log-likelihood test are used to determine the degree of confidence pertaining to the statistical significance of the results of the analysis (Dunning 1993). By conducting this statistical test, researchers are able to move beyond simple descriptions of the data in the corpus.

RESULTS AND DISCUSSION

The following sub-sections present the results of analysis and the discussion of the findings.

THE LEXICAL BUNDLE LIST

A total of 1055 lexical bundles of varying lengths remained on the list after the application of the exclusion criteria. These 1055 bundles amount to a total of 48220 individual cases, which make up 2.19% of one million words in the corpus of this study. As can be expected, 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, or 14% of the total. There are only 10 different five-word lexical bundles in the corpus, representing 0.9% of all bundles. Tables 2, 3 and 4 display the normalised frequencies (per million words) and MI scores of the most frequent three-word, four-word and five-word lexical bundles found in the IBM corpus. It is apparent that the frequency and the length of lexical bundles are inversely related. This observation is in line with the general characteristics of the lexical bundles, that the longer the lexical bundle, the lower is its frequency (Biber et al. 1999; Hyland 2008b; Salazar 2014).

TABLE 2. Top 50 three-word lexical bundles in order of normalised frequency

Rank	Normalised frequency	Mutual information	Three-word lexical bundle
1	452	12.09308	more likely to
2	429	10.52199	in order to
3	413	13.09616	as well as
4	397	9.554226	in terms of
5	370	7.58819	the number of
6	366	10.86638	the relationship between
7	344	6.80119	the level of
8	319	7.420764	the impact of
9	318	13.37095	are more likely
10	296	6.838684	the effect of
11	264	6.636645	the effects of
12	250	8.099408	the importance of
13	248	10.83741	likely to be
14	222	11.65321	the host country
15	220	9.530641	in this study
16	216	11.5612	as a result
17	212	5.923225	the results of
18	209	9.261086	based on the
19	204	7.356016	the role of
20	197	9.932595	are likely to
21	184	8.417365	a number of
22	176	7.479037	on the other
23	176	6.578637	the use of
24	161	8.249528	the presence of
25	160	6.538393	the development of
26	159	9.126263	in addition to
27	155	7.284573	in the host
28	154	6.711911	the context of
29	152	8.462134	of this study
30	151	7.231462	related to the

31	151	4.356357	firms in the
32	149	7.203045	the case of
33	144	9.300939	consistent with the
34	142	8.894184	is likely to
35	140	4.738225	of the firm
36	137	11.66807	is consistent with
37	137	6.328205	the influence of
38	136	8.011634	the likelihood of
39	136	6.145799	the value of
40	134	8.801253	to control for
41	132	13.94162	in other words
42	130	11.63105	we find that
43	130	9.54558	the fact that
44	128	11.70103	in line with
45	128	6.988214	in the same
46	122	12.00551	with respect to
47	121	14.15478	positively related to
48	120	4.784475	the performance of
49	119	12.55809	the dependent variable
50	117	7.991165	the basis of

TABLE 3. Top 50 four-word lexical bundles in order of normalised frequency

Rank	Normalised frequency	Mutual information	Four-word lexical bundle
1	306	18.67982	are more likely to
2	189	16.66825	the extent to which
3	161	19.47854	on the other hand
4	130	11.79915	in the context of
5	120	16.09734	in the host country
6	120	12.22243	in the case of
7	104	14.87262	on the basis of
8	88	8.79913	the results of the
9	87	17.43805	more likely to be
10	81	20.04829	at the same time
11	77	14.81734	as well as the
12	74	20.02879	is positively related to
13	71	11.21537	in terms of the
14	67	16.03404	per cent of the
15	63	11.78639	in the form of
16	62	15.42083	is likely to be
17	60	16.00494	it is important to
18	60	14.50979	as a result of
19	58	12.78551	to the extent that
20	56	17.15052	more likely to have
21	55	15.8141	are likely to be
22	55	15.18342	on the relationship between
23	54	20.00259	a positive relationship between
24	54	18.20775	are less likely to
25	52	9.113387	the size of the
26	50	15.83654	a high level of
27	49	12.4333	the rest of the
28	48	17.24738	a large number of
29	48	15.1708	the degree to which
30	48	14.33804	we find that the
31	46	15.79703	a higher level of
32	46	11.48134	in addition to the
33	44	18.28637	on the one hand
34	44	15.31576	is more likely to
35	44	14.17387	is consistent with the
36	43	10.19876	the nature of the
37	42	21.01111	the liability of foreignness
38	41	16.35266	be more likely to
39	41	14.01293	of the host country
40	40	16.98383	at the country level
41	40	14.5411	with respect to the

42	39	9.138796	of the number of
43	38	18.53694	to take advantage of
44	38	17.99373	a better understanding of
45	38	17.65864	the positive relationship between
46	37	15.34452	the total number of
47	36	16.55024	positively related to the
48	36	14.28781	with regard to the
49	36	14.01535	in line with the
50	35	17.80611	it is possible that

TABLE 4. Top 10 five-word lexical bundles in order of normalised frequency

Rank	Normalised frequency	Mutual information	Five-word lexical bundle
1	55	23.925991	are more likely to be
2	48	24.077653	are more likely to have
3	42	23.417551	firms are more likely to
4	42	18.642718	the extent to which the
5	28	22.771089	is positively related to the
6	28	20.409182	the findings of this study
7	28	17.123934	on the basis of the
8	24	19.158791	the results of this study
9	21	13.313495	in the context of the
10	20	23.631791	they are more likely to

As can be seen, 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. The three-word lexical bundle *more likely to* is an independent bundle which may be arguably subsumed into four-word bundle *are more likely to* and five-word bundle *are more likely to be*. Similarly, the four-word bundle *are more likely to* could also be part of the longer bundle *are more likely to be*. Nevertheless, this shorter three-word bundle *more likely to* which seems to be the fragment of the longer four- and five-word bundles was maintained in this study. This is because the shorter three-word lexical bundle *more likely to* occurs 452 times per million words, much more frequent than the four- and five-word bundles of which it forms part (which occur 306 times and 55 times per million words, respectively). This shows that the three-word lexical bundle *more likely to* has more collocates in its collocational environment. It does not only overlap with the longer bundles *are more likely to* and *are more likely to be*, it also collocates with other words which forms other longer bundles. For instance, *more likely to* is part of other longer bundles such as *is more likely to* (44 times per million words), *more likely to have* (56 times per million words), *are more likely to have* (48 times per million words), and *firms are more likely to* (42 times per million words).

COMPARISON WITH SIMPSON-VLACH AND ELLIS'S (2010) ACADEMIC FORMULAS LIST

To reiterate, there are a total of 1055 types of three- to five-word lexical bundles found in IBM corpus. The top 50 types of lexical bundles of different lengths with their normalised frequencies (per million words) and MI scores are presented in Table 5. It can be seen that all lexical bundles in the top 50 occur more than 100 times per million words. Most of the frequent lexical bundles are in three-word strings, with only 8% of them in 4-word strings. The distinctive four-word bundles are *the extent to which, are more likely to, on the other hand* and *in the context of*.

TABLE 5. Top 50 lexical bundles in IBM in order of normalised frequency

Rank	Normalised frequency	Mutual information	Lexical bundle
1	452	12.09308	more likely to
2	429	10.52199	in order to
3	413	13.09616	as well as
4	397	9.554226	in terms of
5	370	7.58819	the number of
6	366	10.86638	the relationship between
7	344	6.80119	the level of
8	319	7.420764	the impact of
9	318	13.37095	are more likely
10	306	18.67982	are more likely to
11	296	6.838684	the effect of
12	264	6.636645	the effects of
13	250	8.099408	the importance of
14	248	10.83741	likely to be
15	222	11.65321	the host country
16	220	9.530641	in this study
17	216	11.5612	as a result
18	212	5.923225	the results of
19	209	9.261086	based on the
20	204	7.356016	the role of
21	197	9.932595	are likely to
22	189	16.66825	the extent to which
23	184	8.417365	a number of
24	176	7.479037	on the other
25	176	6.578637	the use of
26	161	8.249528	the presence of
27	161	19.47854	on the other hand
28	160	6.538393	the development of
29	159	9.126263	in addition to
30	155	7.284573	in the host
31	154	6.711911	the context of
32	152	8.462134	of this study
33	151	7.231462	related to the
34	151	4.356357	firms in the
35	149	7.203045	the case of
36	144	9.300939	consistent with the
37	142	8.894184	is likely to
38	140	4.738225	of the firm
39	137	11.66807	is consistent with
40	137	6.328205	the influence of
41	136	8.011634	the likelihood of
42	136	6.145799	the value of
43	134	8.801253	to control for
44	132	13.94162	in other words
45	130	11.63105	we find that
46	130	9.54558	the fact that
47	130	11.79915	in the context of
48	128	11.70103	in line with
49	128	6.988214	in the same
50	122	12.00551	with respect to

Table 6 compares the top 50 lexical bundles in IBM corpus with the frequent core academic formulas proposed by Simpson-Vlach and Ellis (2010). The comparison of the results of this study with those of Simpson-Vlach and Ellis (2010) was necessary to determine the specificity of the lexical bundles in this study. To reiterate, Simpson-Vlach and Ellis's list of academic formulas is a cross-disciplinary list of lexical bundles which uses a common-core approach to compile lexical bundles common in various academic disciplines. In contrast, the list of lexical bundles retrieved from IBM corpus is a discipline-specific list of lexical bundles, representing phraseological sequences which are seen specific and

significant in the field of IBM. The comparison between these two lists of lexical bundles is methodologically justifiable as both lists of lexical bundles were retrieved using statistically-driven methods.

TABLE 6. Comparison of lexical bundles with AFL (2010)

Rank	Bundle in this study (IBM)	Bundle in AFL (2010) (Core academic formulas across various disciplines)
1	more likely to	in terms of
2	in order to	the use of
3	as well as	in order to
4	in terms of	as well as
5	the number of	the number of
6	The relationship between	there is a
7	the level of	part of the
8	the impact of	a number of
9	are more likely	the fact that
10	are more likely to	it is not
11	the effect of	there is no
12	the effects of	the case of
13	the importance of	in which the
14	likely to be	in the case
15	the host country	in the case of
16	in this study	based on the
17	as a result	the presence of
18	the results of	due to the
19	based on the	as a result
20	the role of	the role of
21	are likely to	the development of
22	the extent to which	at the same
23	a number of	that there is
24	on the other	likely to be
25	the use of	the effect of
26	the presence of	the basis of
27	on the other hand	such as the
28	the development of	the same time
29	in addition to	with respect to
30	in the host	the effects of
31	the context of	at the same time
32	of this study	it can be
33	related to the	is that the
34	firms in the	on the basis
35	the case of	the importance of
36	consistent with the	in this case
37	is likely to	a variety of
38	of the firm	in relation to
39	is consistent with	can be used
40	the influence of	the context of
41	the likelihood of	in other words
42	the value of	in the same
43	to control for	it may be
44	in other words	a series of
45	we find that	a result of
46	the fact that	is to be
47	in the context of	and in the
48	in line with	the nature of
49	in the same	for example the
50	with respect to	on the basis of

Table 7 presents the list of lexical bundles common in IBM corpus and AFL. Of all the frequent lexical bundles in IBM corpus, 36% of them are seen common in the AFL. This means that 64% of the lexical bundles in IBM are not found in AFL. Also, the statistical measure, the log-likelihood test was performed to study the keyness of the lexical bundles in

IBM and AFL. As keyness is an indicator of specificity, the results of the log-likelihood test show that more than 70% of the shared lexical bundles are more specific to IBM corpus. This indicates that the lexical bundles in IBM corpus are relatively specific as compared with AFL. Also, there are not enough AFL that could cater to the need of learners in the field of IBM. A discipline-specific approach to the teaching and learning of lexical bundles for EAP is seen necessary. This finding is in harmonious with Hyland (2008a) where Hyland demonstrates that there is considerable variation in disciplinary preferences in terms of the types of lexical bundles found in four different academic domains. Over half the lexical bundles in each list did not occur at all in any other discipline in Hyland's (2008a) study, while in this study, more than 60% of the lexical bundles were not found in AFL.

TABLE 7. Lexical bundles common in IBM corpus and AFL

No.	Bundle in the current study (IBM)	Bundle in AFL (Core academic formulas across various disciplines)	Log-likelihood
1	in order to	in order to	+ 60.79
2	as well as	as well as	+ 60.79
3	in terms of	in terms of	+ 60.79
4	the number of	the number of	+ 33.59
5	the effect of	the effect of	+ 126.31
6	the effects of	the effects of	+ 109.17
7	the importance of	the importance of	+ 104.07
8	likely to be	likely to be	+ 69.84
9	as a result	as a result	+ 33.79
10	the role of	the role of	+ 28.82
11	a number of	a number of	- 3.58
12	the use of	the use of	- 27.59
13	the presence of	the presence of	+ 4.11
14	the development of	the development of	+ 7.11
15	the case of	the case of	- 1.78
16	in other words	in other words	+ 11.37
17	the fact that	the fact that	- 22.23
18	in the same	in the same	+ 9.73

Hyland (2008a) proposes that the creation of lists of academic lexical bundles should be discipline-specific oriented as the use of lexical bundles differs by discipline. For instance, Hyland (2008a) reveals that many lexical bundles used in electrical engineering were not found in other academic disciplines, including business studies, applied linguistics and biology. Moreover, electrical engineers were found using the biggest range of different bundles, while biologists employ the fewest bundle types in academic writing. However, Simpson-Vlach and Ellis (2010) argue that they were able to identify lists of lexical bundles that are commonly used in various academic disciplines. The results of this study are in line with those of Hyland (2008a), but are in contrast to Simpson-Vlach and Ellis's (2010). Nonetheless, it should be noted that there are methodological differences between the previous studies and this study.

First, in Hyland's (2008a) study, he investigates the lexical bundles using frequency cut-off threshold, while in Simpson-Vlach and Ellis's (2010) study, both frequency and Mutual Information (MI) cut-off thresholds are set in the corpus tool during the data extraction process. Similar to Simpson-Vlach and Ellis (2010), this study uses both the frequency and MI statistic to retrieve the relevant lexical bundles. The use of frequency and MI statistic in both the present study and in Simpson-Vlach and Ellis's study justifies the comparability of the lists of lexical bundles in both studies. It is worth noting that the use of MI is necessary as MI has been widely known as a good indicator of the association between words. Besides, the sole reliance on frequency count, such as in Hyland (2008a) would most probably overlook some significantly useful expressions with lower frequency count. A

better alternative to the extraction of lexical bundles is to combine the use of frequency and MI statistic, as afforded by corpus tools, such as *Collocate 1.0*.

Second, in Hyland's (2008a) study, only four-word lexical bundles were analysed, while in Simpson-Vlach and Ellis's (2010), three-, four-, and five-word bundles were included in their data set. It is therefore apparent that both the results of this study and those of Simpson-Vlach and Ellis (2010) are relatively more comparable. In view of Simpson-Vlach and Ellis's claim on a common-core approach to the identification and use of lexical bundles for pedagogical purposes, there is a need to verify if there are enough common lexical bundles to facilitate learners with different disciplinary backgrounds. This study is an attempt to explore the issue of specificity with regard to the use of lexical bundles in a specific academic field. The findings of this study indicate that the constructions of academic phraseological sequences need to accord to specific academic needs and purposes.

In sum, in relation to the teaching of academic phrases and expressions, it is convincingly proven that EAP is better approached in a more specific manner. Practitioners in EAP should be provided with added professional training in order to efficiently handle "disciplinary-sensitive repertoire of bundles" (Hyland 2008a, p. 8). EAP instructors are also encouraged to work closely with subject specialists in order to gain a better understanding of subject-related discourse (Hyland 2006).

CONCLUSION

The most frequent lexical bundles in IBM corpus are three-word bundles, including *more likely to*, *in order to*, *as well as*, *in terms of* and *the number of*. The comparisons of lexical bundles in this study with those of Simpson-Vlach and Ellis (2010) indicate that lexical bundles are discipline-specific. The findings of this study have implications on how EAP should be perceived and approached in language classroom. Currently there are debates over the issue of specificity in EAP teaching, influencing both teachers and researchers. Based on the outcome of the analysis, it is suggested that the teaching and learning in EAP should follow a subject- or discipline-specific approach as phraseological sequences such as lexical bundles are highly likely to be markers of disciplines. It is nevertheless never easy to put specificity into practice in EAP classrooms. EAP teachers need to work closely with subject specialists to gain better understanding of the specific language conventions in the respective courses. The collaboration can take various forms, including regular discussions with subject experts. To sum up, there are differing views with regard to the approaches to EAP and this issue remains debatable in the field. It is thus necessary for researchers to continue exploring the various types of phraseological sequences in academic discourse for the sake of further enhancing EAP instructions and syllabuses. The enhancement of EAP syllabuses is crucially important to learners so that they are equipped with the ability to participate in the relevant "academic community" as espoused by Coxhead (2008).

ACKNOWLEDGEMENTS

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

REFERENCES

Ang, L. H., Tan, K. H. & He, M. (2017). A corpus-based collocational analysis of noun premodification types in academic writing. *3L: The Southeast Asian Journal of English Language Studies*. *Vol. 23*(1), 115-131.

Barlow, M. (2004). *Collocate 1.0 software*.

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

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

Braine, G. (1988). A reader reacts (commentary on Ruth Spack's "Initiating ESL students into the academic discourse community: how far should we go?"). *TESOL Quarterly*. *Vol. 22*(4), 702.

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

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*. *34*, 213-238.

Coxhead, A. (2008). Phraseology and English for academic purposes. In F. Meunier & S. Granger (Eds.), *Phraseology in Foreign Language Learning and Teaching* (pp. 149-162). Amsterdam: John Benjamins Publishing Company.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*. *Vol. 19*(1), 61-74.

Durrant, P. (2014). Discipline and level specificity in university students' written vocabulary. *Applied Linguistics*. *Vol. 35*(3), 328-356.

Granger, S. & Paquot, M. (2009). In search of a General academic vocabulary: a corpus-driven study. In K. Katsampoxaki-Hodgetts (Eds.), *Proceedings of the International Conference on L.S.P: Options and practices of LSP practitioners* (pp. 94-108). Heraklion: University of Crete Publications.

Hill, J. (2000). Revisiting priorities: From grammatical failure to collocational success. In M. Lewis (Ed.), *Teaching collocations: Further Development in the Lexical Approach* (pp. 47-69). London: Language Teaching Publications.

Hutchison, T. & Waters, A. (1987). *English for Specific Purposes*. Cambridge: Cambridge University Press.

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

Hyland, K. (2006). *English for Academic Purposes: An Advanced Resource Book*. New York: Routledge.

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

Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*. *Vol. 18*(1), 4-62.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics*. *Vol. 32*, 150-169.

Hyland, K. & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*. *Vol. 41*(2), 235-253.

Jordan, R. (1989). English for Academic Purposes (EAP). *Language Teaching*. *Vol. 22*(3), 153-164.

Jordan, R. (1997). *English for Academic Purposes: A Guide and Resource Book for Teachers*. Cambridge: Cambridge University Press.

Lea, M. & Street, B. (1999). Writing as academic literacies: understanding textual practices in higher education. In C. N. Candlin & K. Hyland (Eds.), *Writing: Texts, Processes and Practices* (pp. 62-81). London: Longman.

Paquot, M. & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics. Vol. 32,* 130-149.

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: Longman.

Salazar, D. (2014). *Lexical Bundles in Native and Non-Native Scientific Writing*. Amsterdam: John Benjamins.

Schutz, N. (2013). How specific is English for Academic Purposes? A look at verbs in business, linguistics and medical research articles. *Language and Computers*. *Vol. 77*(1), 237-257.

Simpson-Vlach, R. & Ellis, N. C. (2010). An academic formulas list (AFL). *Applied Linguistics. Vol. 31*, 487-512.

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

Spack, R. (1988). Initiating ESL students into the academic discourse community: how far should we go? *TESOL Quarterly*. *Vol. 22*(1), 29-52.

Zamel, V. (1993). Questioning academic discourse. *College ESL*. *Vol. 3*, 28-39.