3L: The Southeast Asian Journal of English Language Studies – Vol 25(2): 99 – 112
http://doi.org/10.17576/3L-2019-2502-08

From Lexical Bundles to Lexical Frames: Uncovering the Extent of Phraseological Variation in Academic Writing

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

TAN KIM HUA
Centre of Literacy and Socio-Cultural Transformation,
Universiti Kebangsaan Malaysia

ABSTRACT

The contextual knowledge of a word is closely related to the knowledge of phraseological sequences as words are often used in the phraseological forms, either continuous or discontinuous. Much has been done to examine the continuous phraseological sequences for various purposes. However, studies on phraseology often overlook the potentially useful discontinuous phraseological sequences that allow for more flexible and productive use of language forms. To bridge the gap in phraseology studies, this study therefore employed a corpus-driven approach to analyse the characteristics of a form of discontinuous phraseological sequence, namely lexical frames in a one-million-word corpus of research articles in International Business Management (IBM). The characteristics of lexical frames were observed in four aspects: the degrees of variability and predictability of lexical frames, the structures as well as the variable slot fillers of lexical frames. The corpus tool, Collocate 1.0 was used to extract three- and four-word lexical bundles while kfNgram was used to extract three- and four-word lexical frames from the lexical bundles. The results revealed that three-word lexical frames are more prevalent in IBM. The degree of variability analysis indicated that there are more fixed lexical frames in the category of three-word lexical frames compared to the four-word category. In terms of the degree of predictability, the category of four-word lexical frames contains more predictable lexical frames than the three-word category. Also, most lexical frames are function word frames and the lexical frames are mostly filled up by content words rather than function words. This study contributes to the understanding of phraseological variation in academic writing.

Keywords: corpus-driven; phraseology; discontinuous phraseological sequences; lexical frames; International Business Management (IBM)

INTRODUCTION

With the advances in computer-mediated research methodology, exploring phraseological sequences has become a major area of interest in language and linguistic studies. Linguists have been increasingly intrigued by how words co-occur frequently in language to form regularly used phraseological expressions. One landmark investigation of frequent continuous phraseological sequences is the large-scale study of *lexical bundles* that was published in a contemporary grammar book, *Longman Grammar of Spoken and Written English* (Biber et al. 1999). The study of lexical bundles was based on a corpus analysis of multi-million-word language corpora representing academic prose and conversation. Biber and his colleagues adopted frequency-based approach to identify and compare the structures of lexical bundles in written and spoken registers. In a later study of lexical bundles in university classroom teaching and textbooks, Biber, Conrad and Cortes (2004) developed a functional taxonomy for the categorisation of lexical bundles according to their discourse functions. More recently, authors such as Hyland (2008a) offered another functional classification of lexical bundles for academic genre. The functional classification of the

phraseological expressions is particularly important as it helps learners gain control over the use of lexical bundles in real life contexts.

As there is a growing interest in understanding how continuous phraseological sequences are structured and used in academic discourse, numerous corpus studies have attempted to uncover the role of lexical bundles in characterising academic registers, genres, and disciplines. For instance, Biber et al. (1999) revealed that most lexical bundles are not complete structural units in their corpus of academic writing. These lexical bundles often end in a function word, such as an article or a preposition (e.g. *the context of the*, *as a result of*). The few structurally complete bundles are usually phrases that function as discourse markers (e.g. *in the first place*, *for the first time*). A notable finding by Biber et al. (1999) is closely related to the potentially useful but much neglected discontinuous phraseological sequences. Biber et al. (1999) found that most lexical bundles in academic prose consist of prepositional or nominal elements that co-occur in highly productive frames, such as *the* + * + *of the* + *. The two empty slots represented by the asterisk key * can be filled by many words to make several different lexical bundles (e.g., *the* ***number*** *of the* ***patterns***, *the* ***nature*** *of the* ***business***).

Apart from register-based analysis, Biber et al. (2004) have initiated the exploration of lexical bundles from the genre-based perspective. They compared lexical bundles in textbooks and classroom teaching with those found in their previous research on academic prose and conversation (Biber et al. 1999) by focusing on the structural and functional characteristics of lexical bundles. The comparison revealed that classroom teaching used more discourse organising expressions and stance bundles than conversation. Biber (2006) also made a similar comparison and discovered that classroom teaching used approximately twice as many different lexical bundles as conversation and about four times as many as textbooks. He attributed the extensive use of lexical bundles in classroom teaching to the teachers' needs to provide explanation and elaboration while teaching. Apart from Biber and his colleagues, other scholars (e.g. Cortes 2004, Chen & Baker 2010, Salazar 2014) have forayed into student writing. They concluded that students were found not corresponding to the typical uses of lexical bundles by professional or native writers. In particular, Salazar (2014) proposed specific activities for the teaching of lexical bundles in scientific discourse. Also, in an assertive tone, she described lexical bundles as having distinctive features which can characterise different disciplines. In arguing for the suitability and usability of lexical bundles as markers of disciplines, studies conducted by Ang (2016) and Ang and Tan (2018) showed that the use of lexical bundles differs across disciplines.

With regard to the relevance of phraseology in language learning and teaching, there is a growing awareness of the necessity to incorporate explicit teaching of lexical bundles into language classrooms, for instance in the English for Academic Purposes (EAP) settings. This is evidenced by the empirically derived lists of lexical bundles, namely the Academic Formulas List (AFL) by Simpson-Vlach and Ellis (2010). The AFL was analysed and classified into functional uses based on the framework proposed by Biber et al. (2004) with some modifications. This AFL serves as a good start for placing the teaching and learning of phraseological sequences high on the agenda of language instructors in the field of EAP. Nevertheless, it needs to be reiterated that phraseology is characterised by both continuous and discontinuous phraseological sequences. As such, scholars (e.g. Sinclair 2004, Philip 2008, Biber 2009, Gray & Biber 2013) have reminded us of the importance of all forms of phraseological patterns in language.

Despite the importance of the potentially useful discontinuous phraseological sequences in academic writing, discontinuous phraseological sequences have not received due research attention, as evidenced by the paucity of empirical research published in the field of academic writing. Few scholars have examined discontinuous phraseological sequences under the rubrics of *collocational frameworks* (Renouf & Sinclair 1991),

congrams (Cheng, Greaves & Warren 2006, Cheng et al. 2009), *phrase-frames*/*p-frames* (Biber 2009, Römer 2010, Fuster-Marquez 2014, Garner 2016), and *lexical frames* (Gray & Biber 2013). With the advent of useful corpus analysis tools (e.g. *kfNgram, ConcGram*) which can generate discontinuous phraseological sequences in an automated way, researchers have been able to use corpus-driven approach to study the discontinuous phraseological sequences in a more efficient way.

In an early study of discontinuous phraseological sequences, Renouf and Sinclair (1991) examined frames formed by function words which are termed the collocational frameworks, for example, *a* + * + *of*. They showed evidence that the slot fillers in their collocational frameworks are not random selections. Instead, these slot fillers are seen as belonging to particular semantic groupings. They also asserted that language patterns are not only concerned with lexical words, but also with grammatical words. Language patterns are relatively variable, determined by the elements surrounding them. Biber (2009) expanded on Renouf and Sinclair's (1991) work by introducing a corpus-driven approach to investigate frequent lexical bundles and their variation in conversation and academic writing. He adopted bundles-to-frames approach in identifying the variation of lexical bundles, describing the variation of lexical bundles as phrase frames with slots that are potentially variable (e.g. 1*34, 12*4, *234, 123*). Biber reported that academic writing relies heavily on frames with intervening variable slots and frames are usually formed by function words while variable slots are mostly filled by content words. In contrast, conversational discourse depends more on phrase frames with external variable slots and both the frames and the variable slots are typically filled by function words. Biber insightfully demonstrated that lexical bundles can be approached by looking at the fixedness or variation associated with lexical bundles. Römer's (2010) work on establishing a phraseological profile of a text type included the identification and profiling of phrase frames using bundles-to-frames approach. She concluded that the phraseological profile of a text type is central in determining "the extent of the phraseological tendency of [a] language", which provides "insight into meaning creation in the discourse" (Römer 2010, pp. 309-325). Similar to Biber (2009), Gray and Biber (2013) analysed both lexical bundles and lexical frames in academic prose and conversation. In particular, they examined the characteristics of lexical frames by classifying the structural patterns of lexical frames into several categories. Using direct approach in identifying lexical frames, Gray and Biber (2013) worked on the predictability score of lexical frames and found that lexical frames with low predictability score are usually not associated with any highly frequent lexical bundles, and vice versa. They concluded that the phraseological variation of lexical frames in academic writing is "inherently" associated with grammatical constructions (Gray & Biber 2013, p. 128).

In short, findings of these past studies indicated that there are different degrees and types of variability in the variable slots within the discontinuous phraseological sequences such as phrase frames or lexical frames. As Römer (2010) mentioned, the analysis of phrase frames helps us see to what extent language units allow for variation and this may provide interesting insights into the patterns of phraseological sequences. Also, the phenomenon of variation within the phraseological sequences has not received considerable attention in the literature. In essence, both the continuous and discontinuous phraseological sequences deserve equal attention as both of them characterise the language patterns (Sinclair 2008). Thus, there is a need for research that focuses on discontinuous phraseological sequences in uncovering the phraseological tendency of a language. To bridge the gap in the literature, this study therefore aims to examine the characteristics of discontinuous phraseological sequences, known as lexical frames in journal articles published in the field of International Business Management (IBM). The selection of journal articles in IBM as the corpus for the study was

based on the limited research on the subject and the emerging trend in focusing on business discourse, particularly in Asian context (Bargiela-Chiappini & Zhang 2013).

METHOD

THE CORPUS

The corpus for the study consists of one-million-word tokens, and it includes 138 original research articles, with 59 texts from *Asian Business Management* and 79 from *Journal of International Business Studies*, published from year 2007 to 2013. Both journals are Thomson Reuters-indexed and they achieve satisfactory impact factor yearly. Authors of these two international journals consist of expert writers from various countries.

IDENTIFICATION OF LEXICAL BUNDLES

Following bundles-to-frames approach, the first step of the analysis was to create a list of the most frequent lexical bundles in IBM corpus in order to derive lexical frames. In accordance with Biber et al. (1999), lexical bundle is defined as frequently recurring sequence of words. Biber et al. (1999) proposed that lexical bundles ranging from three to six words are researchable. The study therefore focused on three- and four-word lexical bundles. The steps taken in identifying, retrieving and determining the eligibility of lexical bundles are shown in Figure 1.

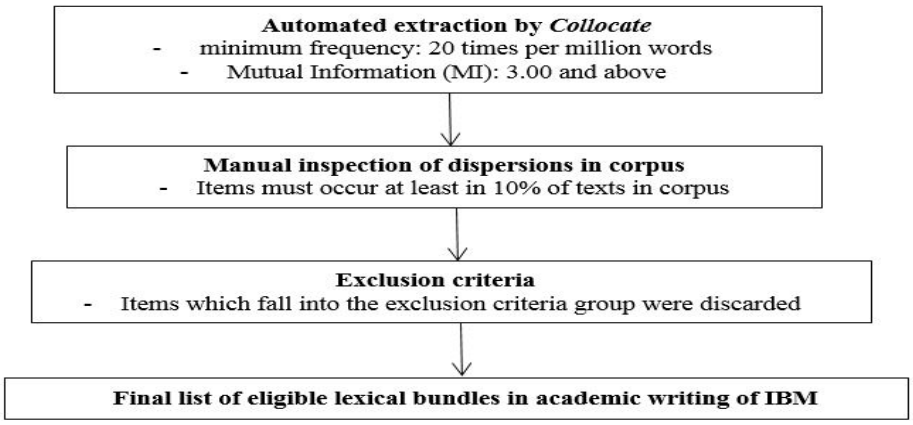

FIGURE 1. Steps in identifying, retrieving and determining the lexical bundles

The software *Collocate 1.0* (Barlow 2004) was used to retrieve lexical bundles automatically by setting the span options. This corpus tool recognises plain text files which end with *.txt* extension. *Collocate 1.0* retrieved lists of n-grams (lexical bundles) by two statistics: frequency and Mutual Information. 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 and 270 four-word sequences. After the extraction by *Collocate 1.0*, the next step was to check the dispersions of the word combinations in corpus. Based on the literature, a phraseological sequence has to occur in three to five texts (Biber & Barbieri 2007) or 10% of texts to avoid idiosyncrasies of particular writers (Hyland 2008b). In the study, it was determined that word combinations which occur in at least 10% of the texts were maintained.

The aim of the study was to make a lexical bundle list that is clear and organised for the purpose of extracting lexical frames from the lexical bundles. The list of lexical bundles

therefore needs to be manually inspected in order to exclude the irrelevant and meaningless word combinations. The exclusion criteria proposed by Salazar (2014) serve as a guide for the study in weeding out irrelevant word combinations. The modified criteria are presented in Table 1. After applying the exclusion criteria, the list of eligible lexical bundles was compiled for the purpose of extracting lexical frames from the relevant lexical bundles.

TABLE 1. Exclusion criteria for lexical bundles

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*

IDENTIFICATION OF LEXICAL FRAMES

The study adopted bundles-to-frames approach in identifying lexical frames. As mentioned, lexical bundles were identified using the software *Collocate 1.0*. After the identification of eligible lexical bundles, the software *kfNgram* (Fletcher 2002) was used to extract the lexical frames automatically from the inventory of lexical bundles. After the identification of lexical frames, only frames with internal variation were maintained as the study intended to look at the internal phraseological variation of phraseological sequences, i.e. lexical bundles. Lexical frames such as '* + *shown*+ *in*' and '*the development* + *is* + *' were excluded as they did not meet the selection criteria of the study.

CHARACTERISTICS OF LEXICAL FRAMES

The distinctive characteristics of lexical frames can be observed in four aspects: the degrees of variability and predictability of lexical frames, the structures and the variable slot fillers of the lexical frames (Biber 2009, Gray & Biber 2013).

In order to study the degree of variability of lexical frames, the variant/p-frame ratio (VPR) measure proposed by Römer (2010, p. 316) was used in this study. The variant/p-frame ratio (VPR) "captures the relation of different words that fill the blank (*) slot in a p-frame to the number of p-frame tokens". The variant refers to the different variable slot fillers of the same lexical frames. For instance, if a lexical frame occurs 500 times and has two variants, it has a VPR of 0.4%. If a lexical frame occurs 500 times but has 300 variants, it has a VPR of 60%. The lower the VPR value, the fewer variants the lexical frame has and that means this particular lexical frame is a rather fixed item, and vice versa. The VPR formula is as follows:

Frequency of variant (filler) type / frequency (token) of lexical frames x 100

Lexical frames are also characterised by their degree of predictability. The degree of predictability is a measure used by Gray and Biber (2013) to determine if a lexical frame has fixed slot filler. The predictability measure is computed by having the token number of the most frequent slot filler divided by the total token number of the lexical frame and then

multiplied by 100. For instance, the lexical frame *more * to* has 500 as its token number (500 times of occurrences). The most frequently found slot filler for this lexical frame is *likely*, which has 452 as its token number (i.e. 452 times the word *likely* fills in the empty slot of *more * to*). The lexical frame *more * to* therefore has a predictability score of 90.4. This suggests that the lexical frame *more * to* is closely associated with a particular high frequency lexical bundle. In this case, the high frequency lexical bundle is *more likely to*. Lexical frames with high predictability scores are always associated with a high frequency lexical bundle, whereas lexical frames with low predictability scores do not have any fixed memberships of frequent slot filler and therefore are not associated with any high frequency lexical bundle. The formula for computing the predictability score is as follows:

frequency of filler / frequency of lexical frames x 100

Besides the degrees of variability and predictability, Gray and Biber (2013) also proposed a broad structural categorisation for the descriptions of the structural correlates of lexical frames. There are three structural categories of lexical frames: frames with other content words, function word frames and verb based frames. For example, *results * that* and *negative * on* were categorised as frames with other content words; lexical frames formed by function words including determiner, preposition and pronoun such as *a * of* and *we * that the* were considered as function word frames; and lexical frames with lexical or auxiliary verb including *measured*, *is* and *be* such as *measured * the*, *be * as* and *is * to* were included in verb based category.

The characteristics of lexical frames were also described in the aspect of variable slot filler. To recap, variable slot fillers are the words that fill in the empty slot within the lexical frames. There are two broad categories of variable slot fillers: content words and function words. These two main categories were further divided into sub-categories, with noun, verb, adjective and adverb grouped under the main category of content words, while determiner, preposition and pronoun were classified as function words fillers. Every lexical frame was assigned to a filler category.

RESULTS AND DISCUSSION

LEXICAL BUNDLES

A total of 1055 lexical bundles remained on the list after the application of the exclusion criteria. The lexical bundle list is largely composed of three-word strings, which account for 85% or 898 of the 1055 target bundles. Tables 2 and 3 display the most frequent three-word and four-word lexical bundles found in the corpus in the descending order of normalised frequency (per million words=pmw).

TABLE 2. Top 20 three-word lexical bundles in order of normalised frequency

Rank	Frequency (pmw)	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

TABLE 3. Top 20 four-word lexical bundles in order of normalised frequency

Rank	Frequency (pmw)	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

CHARACTERISTICS OF LEXICAL FRAMES

Bundles-to-frames approach was adopted to study the phraseological variation within the lexical bundles identified in the study. The inventory of lexical bundles was generated by *kfNgram* tool to sort out the lexical frames. There are three types of lexical frames with internal variability found to be associated with the lexical bundles in the study: 1*3, 1*34 and 12*4. The asterisk mark * indicates variable slot in the lexical frames.

FREQUENCY OF LEXICAL FRAMES

Figures 2 and 3 illustrate the distribution of lexical frames in the study. A total of 125 types and 26781 tokens of lexical frames were retrieved from the relevant lexical bundle inventory. Three-word lexical frames are prevalent in IBM corpus, accounting for almost 77% by type and 87% by token of the lexical frames.

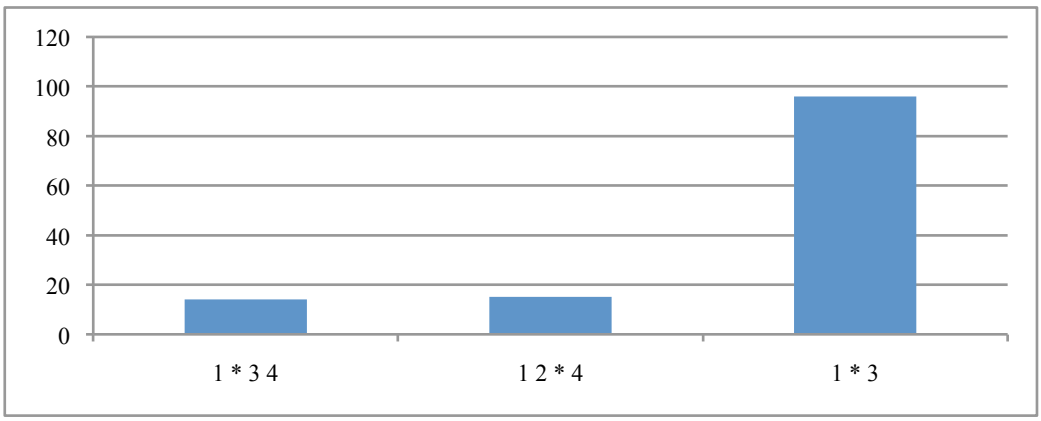

FIGURE 2. Distribution of lexical frames by type

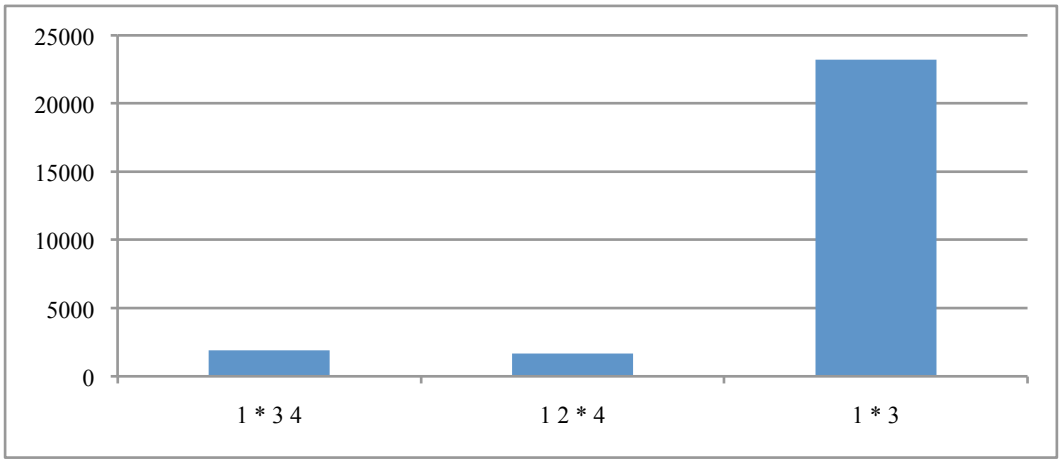

FIGURE 3. Distribution of lexical frames by token

Tables 4 and 5 present instances of three-word and four-word lexical frames found associated with the lexical bundles in the study, respectively. By observing the top five lexical frames, it can be seen that lexical frames are akin to Renouf and Sinclair's (1991) collocational frameworks, where the grammatical words form the frame, leaving an empty slot within the frame for the purpose of studying phraseological tendency and variation. Nevertheless, what differs greatly between lexical frames and collocational frameworks is the ability of lexical frames to identify various types of frames, including frames formed by content words such as *in * markets* (*in foreign markets, in emerging markets*) and *the * environment* (*the institutional environment, the local environment, the business environment*).

TABLE 4. Instances of three-word lexical frames by token order

Rank	Lexical frame	Variant no.	Token no.
1	the * of	105	7529
2	a * of	13	875
3	in * to	5	780
4	to * the	21	753
5	the * between	8	594
6	we * that	8	504
7	more * to	3	500
8	in * of	4	468
9	as * as	2	436
10	are * likely	2	377

TABLE 5. Instances of four-word lexical frames by token order

Rank	Lexical frame	Variant no.	Token no.
1	the * of the	11	414
2	are * likely to	2	360
3	in the * of	5	359
4	the * to which	2	237
5	on the * hand	2	205
6	on the * of	3	162
7	in * host country	2	151
8	in the * country	2	140
9	at the * level	3	101
10	a * level of	2	96

Among the top three-word and four-word lexical frames, 45% of three-word and 30% of four-word lexical frames are partly formed by content words, such as *is * significant*, *influenced * the* and *our results * that*. These examples show that besides the function word based lexical frames (e.g. *an * of*, *a * of the, at the * of*) which are traditionally perceived as the models of frames in language [e.g. the *collocational frameworks* by Renouf & Sinclair (1991)], content word based lexical frames are also able to capture the phraseological variation in academic writing. In other words, the phraseological variation of lexical frames in academic writing are characterised by both grammatical and lexical constructions. To a certain extent, this finding stands in contrast to Gray and Biber's (2013) observation on phraseology in academic writing.

Gray and Biber (2013, p. 128) claimed that the phraseological variation of lexical frames in academic writing is "inherently" associated with grammatical constructions. In the present study, there is corpus evidence that proves that the phraseological variation of lexical frames in academic writing is not "inherently" linked to grammatical constructions only. The lexical frames are formed and characterised by both grammatical and lexical patterning.

The disparity between the finding of the current study and the observation by Gray and Biber (2013) could be due to the methodological differences between the two studies. As mentioned, Gray and Biber (2013) used a direct approach in identifying the lexical frames, while the present study adopted bundles-to-frames approach in retrieving the lexical frames associated with the lexical bundles identified earlier. By using the direct approach, it is not surprising that Gray and Biber were able to retrieve more lexical frames with grammatical constructions as their corpus size is much larger than the one in the current study, which is merely made up of lexical bundle inventory. As the objective of the study was to identify the lexical frames that are associated with the lexical bundles in the study, bundles-to-frames approach was chosen in order to study the internal variability of lexical bundles.

The distinctive characteristics of lexical frames can be observed in four aspects: the degrees of variability and predictability of lexical frames, the structures and the variable slot fillers of the lexical frames (Biber 2009, Gray & Biber 2013).

DEGREE OF VARIABILITY

Tables 6 and 7 present the distributional characteristics of some of the three-word and four-word lexical frames, respectively, showing the variant (type) and token (frequency) numbers as well as VPR score. VPR score is an indication on how variable or fixed a lexical frame is. Gray and Biber (2013) proposed that the degree of variability be divided into three categories, highly variable, variable and fixed. In the study, the degree of variability is determined as follows:

highly variable (VPR>3.5), variable (VPR 2.0-3.5) and fixed (VPR<2.0)

TABLE 6. Instances of three-word lexical frames by descending VPR order

Rank	Lexical frame	Variant no.	Token no.	VPR
1	an * of	3	64	4.69
2	is * significant	3	65	4.62
3	significant * on	2	44	4.55
4	a * impact	2	48	4.17
5	data * the	3	74	4.05
6	is * by	2	50	4.00
7	to * a	3	76	3.95
8	influence * the	2	51	3.92
9	as * by	3	79	3.80
10	to * from	2	53	3.77

TABLE 7. Instances of four-word lexical frames by descending VPR order

Rank	Lexical frame	Variant no.	Token no.	VPR
1	a * of the	2	40	5.00
2	to test * hypotheses	2	40	5.00
3	that the * of	3	62	4.84
4	and the * of	2	42	4.76
5	is * associated with	2	45	4.44
6	our results * that	2	53	3.77
7	to * for the	2	55	3.64
8	of the * of	3	84	3.57
9	the * of this	2	56	3.57
10	at the * of	3	86	3.49

Most lexical frames that constitute the category of three-word lexical frames (1 * 3) are variable lexical frames (46%), followed by fixed lexical frames (35%) and highly variable lexical frames (19%). With regard to the category of four-word lexical frames, most of them are variable lexical frames (45%), followed by highly variable lexical frames (31%) and fixed lexical frames (24%). This shows that there are more fixed lexical frames in the category of three-word lexical frames.

DEGREE OF PREDICTABILITY

Tables 8 and 9 present the distributional characteristics of some of the three-word and four-word lexical frames, respectively, showing the variant (type) and token (frequency) numbers, frequency and type of the most frequent filler for the variable slot and the predictability measure of the lexical frames in the study.

TABLE 8. List of three-word lexical frames by descending predictability measure order

Rank	Lexical frame	Variant no.	Token no.	Filler	Frequency of filler	Predict. score
1	as * as	2	436	well	413	94.72
2	more * to	3	500	likely	452	90.40
3	in * of	4	468	terms	397	84.83
4	are * likely	2	377	more	318	84.35
5	in * host	2	190	the	155	81.58
6	to * extent	2	103	the	82	79.61
7	the * study	2	97	present	76	78.35
8	firms * the	2	195	in	151	77.44
9	the * section	2	81	next	59	72.84
10	we * on	2	83	focus	60	72.29

TABLE 9. List of four-word lexical frames by descending predictability measure order

Rank	Lexical frame	Variant no.	Token no.	filler	Freq of filler	Predictability score
1	in the * country	2	140	host	120	85.71
2	are * likely to	2	360	more	306	85.00
3	the * to which	2	237	extent	189	79.75
4	in * host country	2	151	the	120	79.47
5	on the * hand	2	205	other	161	78.54
6	is * related to	2	96	positively	74	77.08
7	it is * to	2	81	important	60	74.07
8	as a * of	2	84	result	60	71.43
9	a * relationship between	2	78	positive	54	69.23
10	a high * of	2	75	level	50	66.67

The predictability measure is an indicator of the degree of association between the variable slot filler and the lexical frames. Frames that have a low predictability scores do not have frequent and fixed slot filler. They are therefore not associated with the frequent lexical bundles. On the other hand, frames with high predictability scores usually have fixed membership of slot fillers. They are therefore directly associated with the frequent lexical bundles. In the study, the degree of predictability is determined as follows:

highly predictable (predictability score>61), predictable (predictability score 31-60) and unpredictable (predictability score <30)

Most lexical frames that constitute the category of three-word lexical frames (1 * 3) are predictable lexical frames (63%), followed by highly predictable lexical frames (30%) and unpredictable lexical frames (7%). With regard to the category of four-word lexical frames, there are equal numbers of the lexical frames in both the categories of predictable lexical frames (48%) and highly predictable lexical frames (48%). The unpredictable lexical frames only constitute 4% of the category of four-word lexical frames. Overall, three-word lexical frames contain more predictable lexical frames than the four-word lexical frames, while four-word lexical frames contain more highly predictable lexical frames than three-word lexical frames.

STRUCTURES OF LEXICAL FRAMES

Gray and Biber (2013) proposed a broad structural category for the descriptions of the structural correlates of lexical frames. Tables 10, 11 and 12 present instances of lexical frames in the categories of frames with other content words, function word frames and verb based frames, respectively while Table 13 shows the broad structural categories of the lexical frames.

TABLE 10. Instances of lexical frames with other content words

Lexical frames	Variant no.	Token no.
are * likely	2	377
the * country	3	340
in * study	3	339
of * study	3	233
high * of	3	216
firms * the	2	195
of * firm	2	194
in * host	2	190
in * markets	2	188
results * that	3	162

TABLE 11. Instances of function word lexical frames

Lexical frames	Variant no.	Token no.
the * of	105	7529
a * of	13	875
in * to	5	780
to * the	21	753
the * between	8	594
we * that	8	504
in * of	4	468
as * as	2	436
are * to	5	366
the * that	8	363

TABLE 12. Instances of verb based lexical frames

Lexical frames	Variant no.	Token no.
be * to	6	220
is * with	2	200
this * is	5	179
is * to	5	265
have * that	4	121
are * with	2	116
measured * the	2	111
were * to	3	84
are * in	3	81
been * to	2	67

TABLE 13. Broad structural categories of lexical frames

	Variant no.	Token no.
Frames with other content words	128 (27%)	7103 (27%)
Function word frames	299 (63%)	17444 (65%)
Verb based frames	50 (10%)	2234 (8%)

As shown in Table 13 above, it is evident that function word frames are prevalent in IBM corpus, while the lexical frames with other content words are the second most common lexical frames, followed by verb based frames. Again, these results are different from those of Gray and Biber (2013), where they found that academic writing depends more on function word frames and verb based frames. Frames made up of other content words are rarely found in academic writing.

VARIABLE SLOT FILLERS

Variable slot fillers are the words that fill in the empty slot within the lexical frames. Table 14 presents the distribution of the slot filler by word class.

TABLE 14. Word class of variable slot fillers in lexical frames

	Content words		Function words	
	Type	Token	Type	Token
Noun	245 (52%)	14889 (56%)		
Verb	102 (21%)	3891 (15%)		
Adjective	64 (13%)	4808 (18%)		
Adverb	5 (1%)	165 (1%)		
Determiner			29 (6%)	1687 (6%)
Preposition			22 (5%)	933 (3%)
Pronoun			10 (2%)	408 (1%)
Total	416 (87%)	23753 (90%)	61 (13%)	3028 (10%)

Most of the variable slot fillers are content words (nouns, verbs, adjectives or adverbs). In a register-based study of lexical frames, Biber (2009) discovered that content words are typically found as variable elements of lexical frames in academic prose, whereas conversation prefers function words. In this respect, the results of the study are in line with Biber's finding on academic prose.

In short, the analysis of lexical frames shows that lexical frames are observable phraseological patterns of word sequences in academic writing. They are characterised mainly by their degrees of variability and predictability, structures and the variable slot fillers. Results from the variability analysis showed that besides having the traditional function word frames, there are content word based frames which are variable in the IBM academic writing. It could be concluded that lexical frames in IBM corpus are formed and characterised by both grammatical and lexical patterning.

CONCLUSION

The results of the study are likely to have considerable implications for researchers working on phraseology. In the literature, research on phraseology has always focused on continuous phraseological items such as lexical bundles and collocations. Discontinuous phraseological sequences did not receive much attention in the past, even though the concept of discontinuous phraseological sequences was proposed by Renouf and Sinclair (1991) back in year 1991.

This study has made a number of findings which clarify the stereotypical perception about phraseology whereby phraseology had long been perceived as fixed expressions. This perception had led to other forms of phraseology being ignored (Sinclair 2008) for a long time. By analysing both continuous and discontinuous phraseological sequences, we are able to understand the actual phraseological tendency in academic language and to what extent the language allows for variation.

The study also has pedagogical implications on language teaching. Lexical frames with high predictability scores are pedagogically valuable and meaningful. Language teachers can expose learners to another perspective of phraseological variation using these lexical frames that are always associated with particular lexical bundles. Lastly, the study has focused on discontinuous phraseological sequences that possess internal variations. In Biber's (2009) study, the discontinuous phraseological sequences also show interesting external variations, i.e. empty slot on the first or last position. Future research may attempt to study this pattern as there is truly an emerging need to understand more about the patterns in phraseological variation, as asserted by Sinclair (2008).

ACKNOWLEDGEMENTS

This work was supported by Universiti Sains Malaysia Short Term Grant (304/PHUMANITI/6315044 and Universiti Kebangsaan Malaysia Research Universiti Grant coded GUP-2017-079).

REFERENCES

Ang, L. H. (2016). *Phraseological profile of journal articles in International Business Management: A corpus-driven analysis*. (Unpublished doctoral dissertation). Universiti Kebangsaan Malaysia.

Ang, L. H. & Tan, K. H. (2018). Specificity in English for Academic Purposes (EAP): A corpus analysis of lexical bundles in Academic Writing. *3L: The Southeast Asian Journal of English Language Studies*. *24*(2), 82-94.

Bargiela-Chiappini, F., & Zhang, Z. C. (2013). Business English. In B. Paltridge, & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 193-211). Oxford: Wiley-Blackwell.

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

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics. 14*(3), 275–311.

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

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

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

Cheng, W., Greaves, C. & Warren, M. (2006). *From n-gram to skipgram to concgram. International Journal of Corpus Linguistics. 11*(4), 411-433.

Cheng, W., Greaves, C., Sinclair, J. & Warren, M. (2009). Uncovering the extent of the phraseological tendency: Towards a systematic analysis of concgrams. *Applied Linguistics*. *30*(2), 236-252.

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

Fletcher, W. H. (2002). *KfNgram software*. Annapolis. MD: USNA.

Fuster-Marquez, M. (2014). Lexical bundles and phrase frames in the language of hotel websites. *English Text Construction. 7*(1), 84-121.

Garner, J. R. (2016). A phrase-frame approach to investigating phraseology in learner writing across proficiency levels. *International Journal of Learner Corpus Research. 2*(1), 31-68.

Gray, B. & Biber, D. (2013). Lexical frames in academic prose and conversation. *International Journal of Corpus Linguistics. 18*(1), 109-135.

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

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

Philip, G. (2008). Reassessing the canon: 'fixed phrases' in general reference corpora. In S. Granger & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspectives* (pp. 95-108). Amsterdam: John Benjamins.

Renouf, A. J. & Sinclair, J. (1991). Collocational frameworks in English. In K. Ajimer, & B. Altenberg (Eds.), *English corpus linguistics. Studies in honour of Jan Svartvik* (pp. 128-143). Harlow: Longman.

Römer, U. (2010). Establishing the phraseological profile of a text type: The construction of meaning in academic book reviews. *English Text Construction*. *3*(1), 95-119.

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

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

Sinclair, J. (2004). *Trust the text*. London: Routledge.

Sinclair, J. (2008). The phrase, the whole phrase and nothing but the phrase. In S. Granger, & F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 407-410). Amsterdam: John Benjamins.