A Corpus-driven Analysis of Lexical Frames in Academic Writing

Ang Leng Hong
School of Humanities, Universiti Sains Malaysia

lenghong@usm.my

Introduction

In recent years, there is a growing interest in understanding how multi-word sequences, particularly the continuous ones, are structured and used in academic discourse. For instance, in analysing academic prose, Biber et al. (1999) revealed that most continuous multi-word sequences, i.e. 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. *as a result of, the context of the*). 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 multi-word sequences. They 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 different lexical bundles (e.g., *the* ***number*** *of the* ***patterns***, *the* ***nature*** *of the* ***business***).

Research on multi-word sequences in academic registers have shown the relevance of multi-word sequences in academic writing. Thus, there is a growing awareness of the necessity of incorporating explicit teaching of multi-word sequences such as lexical bundles into language classrooms (Biber, Conrad & Cortes, 2004; Hyland, 2008; Salazar, 2014; Ang & Tan, 2018). Nevertheless, researchers in the field have yet to give due research attention to another type of multi-word sequences, the discontinuous ones. As reminded by the scholars in the field, language is characterised by both continuous and discontinuous multi-word sequences and they are equally important language patterns in language (Sinclair, 2004; Philip, 2008; Biber, 2009; Gray & Biber, 2013).

In an early study of discontinuous multi-word 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 belonging to particular semantic groupings. With the advances in corpus linguistics in recent years, Biber (2009) began to investigate frequent lexical bundles and their variation in conversation and academic writing and he described the variation of lexical bundles as phrase frames with slots that are potentially variable (e.g. 1*34, 12*4, *234, 123*). Biber found 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. Biber insightfully demonstrated that lexical bundles can be approached by looking at the fixedness or variation associated with lexical bundles. Similar to Biber (2009), Gray and Biber (2013) analysed both lexical bundles and the discontinuous multi-word sequences, i.e., lexical frames in academic prose and conversation. They 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, 2013:128).

Findings of these past studies indicated that there are different degrees and types of variability in the variable slots within the discontinuous multi-word 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 multi-word sequences. Also, the phenomenon of variation within the multi-word sequences has not received considerable attention in the literature. There is a need for research that focuses on discontinuous multi-word sequences in uncovering the phraseological tendency of the language. To bridge the gap in the literature, this study therefore aims to examine the characteristics of discontinuous multi-word sequences, known as lexical frames in journal articles published in the field of International Business Management (IBM).

Methodology

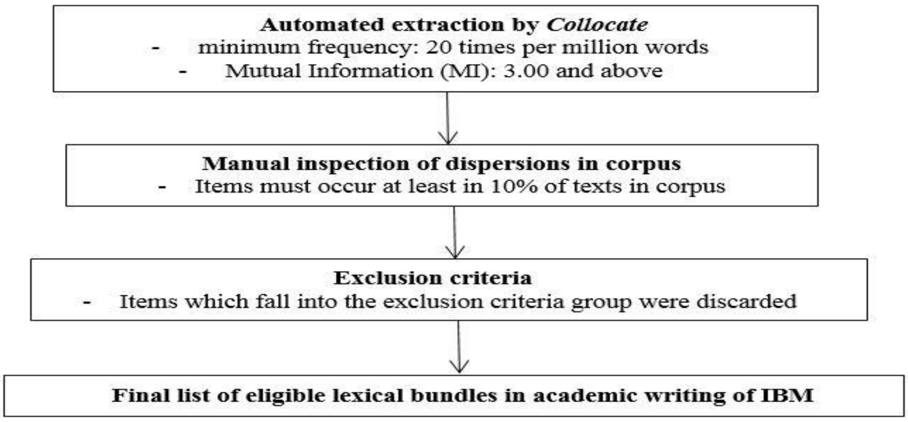
The corpus

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 bundle-to-frame 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. The study focused on three- and four-word lexical bundles. Following the literature, 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


Automated extraction by *Collocate*
- minimum frequency: 20 times per million words
- Mutual Information (MI): 3.00 and above
Manual inspection of dispersions in corpus
- Items must occur at least in 10% of texts in corpus
Exclusion criteria
- Items which fall into the exclusion criteria group were discarded
Final list of eligible lexical bundles in academic writing of IBM

Identification of lexical frames

The study adopted bundle-to-frame 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 retained as the study intended to look at the internal phraseological variation of multi-word sequences, i.e. lexical bundles.

Characteristics of lexical frames

The distinctive characteristics of lexical frames can be observed in two main aspects: the degrees of variability and predictability of 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: 316) was used in this study. 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 was a measure used by Gray and Biber (2013) to determine if a lexical frame has fixed slot filler. 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

Results

Lexical bundles

A total of 1055 lexical bundles of varying lengths 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. Examples of lexical bundles include *more likely to*, *the extent to which*, *in the context of* and *in terms of the*.

Characteristics of lexical frames

Bundle-to-frame 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* software to sort out the lexical frames. There are three types of lexical frames with internal variability found 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. 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.

Degree of variability

Tables 1 and 2 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 of 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 1: 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 2: 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 3 and 4 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 3: 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 4. 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

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.

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 multi-word sequences such as lexical bundles and collocations. Discontinuous multi-word sequences did not receive much attention in the past, even though the concept of discontinuous multi-word 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 multi-word sequences whereby multi-word sequences had long been perceived as fixed expressions. This perception led to other forms of multi-word sequences being ignored (Sinclair 2008) for long time. By analysing both continuous and discontinuous multi-word 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 instructors can expose learners to another perspective of phraseological variation using these lexical frames that are always associated with particular lexical bundles in EAP teaching.

References

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.

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

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.

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

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

Hyland, K. (2008). 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 multi-word 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.

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.