Frequency and structure of Indonesian lexical bundles on academic prose in legal studies: A driven-corpus approach

Adi Budiwiyanto[1] and Totok Suhardijanto[2]
{[1]adibudiwiyanto@gmail.com, [2]suhardiyanto@gmail.com}

[12]Universitas Indonesia

Abstract Lexical bundles are important elements in academic writing. Lexical bundles have been defined as combination of three or more words which are identified in a corpus of natural language by means of corpus analysis program. Every register has their own characteristics. This study aims to describe the frequency and structure of lndonesian lexical bundles on academic prose in legal studies. In this study, the identification of lexical bundles uses a frequency cutoff of 40 per million word that occurs at least within 5 different texts using AntConc computer program. The corpus compiled comprises 2,054,312 words, taken from four genres, i.e. undergraduate thesis, thesis, dissertation and journal article of legal studies. The results reveal that there are 475 lexical bundles, ranging from three to seven words, and mostly composed in three-word combination, such as *oleh karena itu* 'because of that', dalam hal ini 'in this case' and *yang dilakukan oleh* 'which is done by'. Among the four types of academic prose, journal article has the highest number of lexical bundles. In addition to the structural classification, it is found that the corpus has around 68,25 percent incomplete structures of phrase and clause. The patterns of *yang* + verb + prepositional phrase fragments (e.g., *yang diatur dalam, yang ditetapkan oleh*, *yang berhubungan dengan*); noun phrase + *yang*-relative clause fragments (e.g., *alat bukti yang, aturan hukum yang, hukum pidana yang, ketentuan hukum yang*); preposition + noun phrase fragments (*dalam jangka waktu, dalam ketentuan pasal, dalam undang-undang nomor*) are commonly used in this register.

Keywords: lexical bundle, academic prose, corpus, frequency, structure

1. Introduction

One of the most productive topics in language formulaic research in the past several decades has been the area of lexical bundles. The term *lexical bundle* was first used by Biber, Johansson, Leech, Conrad and Finnegan in an English grammar book entitled *Longman Grammar of Spoken and Written English* (1999). The book was compiled using a corpus-based approach. Ref. [1] defines the lexical bundle as a recurring sequence of three or more words in a register, such as *I don't want to*, *I thought that was*, *the nature of the* and *as a result*

BASA 2019, September 20-21, Surakarta, Indonesia
Copyright © 2020 EAI
DOI 10.4108/eai.20-9-2019.2296703

of. These bundles can be regarded as extended collocations, i.e. bundles of words that show a statistical tendency to co-occur.

Corpus studies that have been conducted recently, especially in English, indicate that lexical bundles pervasively appear in various texts in academic registers [1], [2], [3], [4], [5] [6] [7]. In general, lexical bundles are not idiomatic in terms of meaning and not complete in terms of structure. Ref. [4] found that only 15% of the lexical bundles in the conversation register are in the form of complete phrases or clauses, while in the academic prose register there are less than 5% in the form of complete structural units. Although the meaning of the lexical bundles is not idiomatic and the structure is incomplete, they provide a kind of "parent" for larger phrases and clauses. In general, lexical bundles can be used 1) as a means of providing a frame for interpretation of the subsequent proposition, 2) to organize the flow of ideas in discourse and 3) to identify an entity or point out some specific quality of an entity [8]. The functions and meanings expressed by the lexical bundles differ in every register and academic discipline and they depend on their specific objectives.

The comparison of registers conducted by [1] on the study of lexical bundles in English has shown the extent to which lexical bundles are used, both in conversation and academic writing. They show that there are lexical bundles that appear pervasively in written academic register and have certain features in academic writing. For example, in written academic register, lexical bundles often appear in the pattern of noun phrases + fragments of phrases-of (e.g., *the base of the*, *the structure of the*) and the pattern of fragment *it* (e.g., *it is possible to*, *it should be noted that*), while in conversation register the bundles are marked by the pattern of pronouns + verbal phrases (e.g., *I don't think so*, *I said to him*) and auxiliary verbs + active verb fragments (e.g., *going to have a* and *don't worry about it*). In addition, [9]states that the lexical bundles contained in academic writing are generally patterned as prepositions + nominal phrase fragments, nominal phrases + fragment phrases-*of* (e.g., *the base of the*, *the structure of the*) or anticipatory *it* fragments (e.g., *it is possible to*, *it should be noted that*). Those structures represent about 70 percent of the four-word cluster pattern in academic discourse and are rarely found in conversation.

Ref. [10] note that lexical bundles have three main beneficial effects in academic writing: 1) they offer ready-made sets of words to use as a partial foundation for crafting academic prose; 2 they facilitate and represent fluent language use and signal that a writer is a “member” of a discourse community; and 3) they represent register-specific ways of expressing particular meanings. The use of bundles also helps guide readers through text, by signaling linkage of ideas, the writer stance, or the attitude implicit in prose [8].

Researches relating to lexical bundles in Indonesian, to the best of the author's knowledge, have not much been conducted. Therefore, this research study lexical bundles in Indonesian language, especially written academic register. The authors presume that there are special characteristics of lexical bundles in Indonesian given the different characteristics with English. The characteristics studied in this paper include the frequency of lexical bundles and their distribution in four genre, namely undergraduate theses, postgraduate theses, dissertations and journal articles. In addition, this paper also describes the structural classification and grammatical pattern of lexical bundles. This research focuses on academic writing in legal studies.

2. Research Method

This research uses a corpus-driven approach. It is an inductive approach. Inductive reasoning starts from specific observations to broader generalizations and theories. Therefore, this approach is also called bottom-up approach [11]. The researcher begins with specific measurements and observations to identify patterns and order. When patterns are found, researchers formulate tentative hypotheses that can be traced further and may develop into general conclusions.

The corpus prepared for this study is a written academic register corpus. The genre includes undergraduate theses, postgraduate theses, dissertations and journal articles. The text samples in this study are taken from University of Indonesia, University of Sumatera Utara and University of Hasanuddin. Meanwhile, research articles are taken from various journals of legal studiespublished by universities or research institutions that are nationally indexed.

To obtain a representative corpus, the academic discourse texts used in this study were selected by stratified random sampling. The texts used are published from 2010 to 2018. Heterogeneous topics in every genre are mainly concerned. Meanwhile, article selection is based on heterogeneous volume and journal publishers. The author's name can only appear once to avoid idiosyncrasy. The following is recapitulation of the amount of text and words in this research.

Table 1. Composition of legal studies corpus

No.	Genre	Number of			Average of text length
		Type	Token	Text	
1	undergraduate thesis	19.584	516.252	28	18.438
2	thesis	18.990	520.083	27	19.262
3	dissertation	19.035	515.889	10	51.589
4	journal article	20.246	502.088	133	3.775
5	all-over	42.309	2.054.312	198	12.671

3. Results And Discussion

In identifying the lexical bundles, this study uses a threshold frequency of 40 million per word (pmw) and a range (text frequency) of at least five different texts. The corpus used in this study consists of 2,054,312 words. If this study uses a cutoff 40 pmw, it means that a series of words must occur at least 80 times to be categorized as lexical bundle. Based on the predetermined identification parameters, by using [12] computer program, it was identified that the corpus of legal studies has a number of lexical bundles that range from three words to seven words. The three-word bundles is the highest in number, namely 400 bundle (84.21%), while the lowest is seven-word bundles with only 2 bundles (0.42%). Based on the calculation, it can be inferred that the smaller the number of words that make up the lexical bundle, the greater the lexical bundle in number.

The co-occurrence of three-word bundles is seven times as many as the number of four-word bundles, while the co-occurrence of four-word bundles is almost four times as many as the five-word bundles. When compared to the [5]'s study, the ratio is quite significantly different; the appearance of four-word bundles is ten times as many as the five-word bundles and that becomes the consideration in selecting the four-word bundles as the focus of his analysis. Meanwhile, this research focuses on three-word bundles for the analysis of structural classification. The details of lexical bundles based on number, percentage and ratio can be seen in Table 2.

Table 2. Lexical bundles based on word length

No.	Lexical Bundles	Number	Percentage	Ratio
1	3-word	400	84,21	200
2	4-word	53	11,16	26,5
3	5-word	15	3,16	7,5
4	6-word	5	1,05	2,5
5	7-word	2	0,42	1
	Total	**475**	**100**	

The legal studies corpus is divided into four sub-corpora: 1) undergraduate thesis, 2) thesis, 3) dissertation and 4) journal article. Each subcorpus consists of approximately 500 thousand words on average. Based on the analysis by [12], it reveals that lexical bundles are mostly found in journal article by 509 bundles, while the lowest is in dissertation, by 292 bundles. The lexical bundles in undergraduate theses and theses are not much different in terms of number, but both are significantly different compared to dissertation. This demonstrates a tendency that at the highest level of education, lexical bundles are less productively used. Journal article belongs to research academic discourse [5], a published academic writings, where the publications are usually through the process of language reviewing and editing. It is different from the other three texts which belong to student academic discourse, the unpublished one. The percentage of lexical bundle usage can be seen in the following table.

Table 3. Distribution of sub-corpus based on genre

No.	Sub-corpus	Lexical Bundle					Total	%
		3-word	4-word	5-word	6-word	7-word		
1	Skripsi	368	46	10	2	0	426	25,27
2	Theses	398	57	4	0	0	459	27,22
3	Dissertation	269	17	3	2	1	292	17,32
4	Journal article	437	60	9	2	1	509	30,19
						Total	**1686**	**100**

The central topic in legal studies mainly relate to corruption and law enforcement apparatus. This can be seen from the lexical bundles that emerged, namely *tindak pidana korupsi* and *hak asasi manusia* (see Table 4). The table also shows that there are three lexical bundles with high frequencies (more than 800), i.e., *oleh karena itu* and *dalam hal ini* and high ranges (more than 140), i.e., *yang dilakukan oleh*. The former has complete structures and the latter are incomplete. Bundle *oleh karena itu* is usually used as a transitional marker in sentences, whereas bundle *dalam hal ini* is usually used for framing. Different from the two previous bundles, bundle *yang dilakukan oleh* indicates events, actions or methods relating to the research conducted. In addition, there are other lexical bundles in the form of fragments, such as *yang dilakukan oleh*, *yang berkaitan dengan*, *yang diatur dalam* and *diatur dalam pasal*. The following is a list of the top 10 lexical bundles.

Tabel 4. The top 20 lexical bundles in legal studies

No.	Lexical Bundles	Frequency	Range
1	oleh karena itu	972	154
2	dalam hal ini	913	146
3	yang dilakukan oleh	823	143
4	tindak pidana korupsi	664	34
5	negara republik indonesia	603	94
6	aparat penegak hukum	565	67
7	diatur dalam pasal	563	114
8	yang berkaitan dengan	535	131
9	sistem peradilan pidana	495	43
10	tahun 2009 tentang	493	59

Ref. [3] states that one of the characteristics of lexical bundles is incomplete structural unit. Its number reaches 95 percent in academic prose register. Although the structure is incomplete, lexical bundles have strong grammatical correlation. Ref. [1] classifies lexical bundles into several types of basic structures and these structures vary according to the register. In academic writing most lexical bundles are noun phrase fragments and of prepositional phrase fragments.

In this paper, the structural analysis of lexical bundles is carried out only on three-word bundles that consist of 400 bundles because the three-word clusters are more varied in structure. The result of analysis indicates that there are complete structures and incomplete structures in the corpus. Incomplete structures can be found in the form of clauses and phrases. Of that composition, 68.25 percent are bundles with incomplete structure, while the rest, 31.75 percent, are bundles with complete structures. When they are compared to the composition of incomplete structures in academic prose in English, there is significant difference even though incomplete structures still dominating in number. The structural classification of lexical bundles in legal studies can be seen in Table 5.

The grammatical pattern can be divided into four groups, namely 1) lexical bundles that incorporate noun phrase fragments, 2) lexical bundles that incorporate verb phrase fragments, 3) lexical bundles that incorporate prepositional phrase fragments and 4) lexical bundles that incorporate dependent clause fragments. In the corpus, it is also found grammatical patterns in the form of adverbal fragments, but the number is not significant.

There are several grammatical patterns that are salient in this register [14]. Pattern of noun phrase + *yang*-relative clause fragments (e.g., *alat bukti yang, aturan hukum yang, hukum pidana yang, ketentuan hukum yang*) is widely used in bundles that incorporate noun phrase fragments. Pattern of verb phrase + prepositional phrase fragments (e.g., *bertanggung jawab atas*, *telah diatur dalam*, *tidak bertentangan dengan*, *tidak terlepas dari*) and passive verb + prepositional phrase fragments (e.g., *diancam dengan pidana*, *diatur dalam pasal*, *dimaksud pada ayat*) are frequently used in bundles that incorporate verbal phrase fragments. In addition, patterns of *yang* + verb + PP fragments (e.g., *yang diatur dalam, yang ditetapkan oleh*, *yang berhubungan dengan*) is often found in bundles that incorporate dependent clause fragments. The last, pattern of preposition + noun phrase fragments (*dalam jangka waktu, dalam ketentuan pasal, dalam undang-undang nomor*) frequently occurs.

Table 5. Structural classification of lexical bundle

Structural Classification	phrase	complete	NP	aparat penegak hukum, bahan hukum primer, hak asasi manusia, hak dan kewajiban, penggugat dan tergugat
			PP	antara para pihak, dalam hal ini, dengan kata lain, pada saat itu, dengan pidana penjara, di samping itu
			VP	hanya dapat dilakukan
			AdvP	sama sekali tidak
		incomplete	NP fragment	alat bukti yang, aturan hukum yang, para pihak yang, penyelesaian sengketa melalui, perlindungan hukum bagi
			PP fragment	sebagaimana diatur dalam, sebagaimana dimaksud dalam, dalam kaitannya dengan, sebagai bagian dari
			VP fragment	dapat dilakukan dengan, dapat dilihat dari, telah diatur dalam, tidak bertentangan dengan, tidak sesuai dengan, tidak terlepas dari, diancam dengan pidana, diatur dalam pasal
			AdvP fragment	secara sah dan (meyakinkan)
	clause	complet e	Clause fragment	yang ada dalam, yang diatur dalam, yang tercantum dalam,yang diajukan oleh, yang ditetapkan oleh, yang berkaitan dengan, yang bertentangan dengan, yang bertujuan untuk

4 Conclusion

The results above indicate the characteristics of Indonesian lexical bundles in written academic register in legal studies. The registerhas a big number of lexical bundles, ranging from three words to seven words which are dominated by three words, namely 400 bundles (84.21%). In terms of genre distribution, lexical bundles are mostly found in journal articles by 509 bundles (30%), while the lowest is in dissertation by 292 bundles (17.32%). In terms of structure, the bundles consist of incomplete and complete structures. Incomplete structures can be found in the form of clauses and phrases, whereas complete structures can only be found in the form of phrases. This incomplete structures dominate the lexical bundles in legal studies by 68.25 percent.

References

[1] D. Biber, S. Johansson, G. Leech, S. Conrad and E. Finegan, Longman Grammar of Spoken and Written English, Harlow: Longman, 1999.

[2] D. Biber, S. Conrad and V. Cortes, "If you look at…: Lexical bundles in university teaching and textbooks," Applied Linguistics, 25 (3), p. 371–405, 2004.

[3] D. Biber, University language: A corpus-based study of spoken and written registers, Philadelphia: John Benjamins, 2006.

[4] D. Biber and F. Barbieri, "Lexical bundles in university spoken and written registers," English for Specific Purposes, 26 (3), p. 263–286., 2007.

[5] K. Hyland, Bundles in academic discourse, Cambridge: Cambridge University Press, 2012, p. 150–169.

[6] D. Salazar, Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching, Amsterdam ; Philadelphia: John Benjamins Publishing Company, 2014.

[7] P. Byrd and A. Coxhead, "On the other hand: Lexical bundles in academic writing and in the teaching of EAP," University of Sydney Papers in TESOL, vol. 5, no. 5, pp. 31-64, 2010.

[8] D. Wood, Fundamentals of formulaic language: an introduction, London: Bloomsbury Publishing, 2015.

[9] K. Hyland, "As can be seen: Lexical bundles and disciplinary variation.," English for Specific Purposes, 27, p. 4–21, 2008.

[10] A. Coxhead and P. Byrd, "Preparing writing teachers to teach the vocabulary and grammar of academic prose," Journal of Second Language Writing, pp. 129--147, 2007.

[11] W. Cheng, Exploring corpus linguistics: Language in action, London: Routledge, 2012.

[12] L. Anthony, AntConc (Version 3.5.7)[Computer Software]., Tokyo, Japan: Waseda, 2018.

[13] K. Hyland, Academic Discourse: English in Global Context, London, New York: Continuum, 2009.

[14] K. Saddhono and M. Rohmadi, "A Sociolinguistics Study on the Use of the Javanese Language in the Learning Process in Primary Schools in Surakarta, Central Java, Indonesia." Int. Edu. Stu., vol. 7 no.6 pp 25-30, 2014