INVESTIGATING LEXICAL BUNDLES IN THE CORPORA OF ENGLISH AND INDONESIAN RESEARCH ARTICLES WITH THE SKETCH ENGINE

MENYELISIK SIMPUL LEKSIKAL DALAM KORPUS ARTIKEL ILMIAH BERBAHASA INGGRIS DAN INDONESIA DENGAN PERANGKAT LUNAK SKETCH ENGINE

Susi Yuliawati[1], Dian Ekawati[2], Ratna Erika Mawarrani[3]

English Studies Program, Universitas Padjadjaran, Bandung[1,3]
German Studies Program, Universitas Padjadjaran, Bandung[2]

susi.yuliawati@unpad.ac.id[1]

ABSTRACT

The low publication rate of Indonesian researchers in reputable international journals, particularly in arts and humanities, is caused, among others, by difficulties they faced in producing precise expository texts in English, which are different from texts in Indonesian. The present study examines lexical bundles in the corpora of English and Indonesian research articles (RA) on literature and linguistics to describe the similarities and differences of conventionalized phraseology in the scientific genre of English and Indonesian by using corpus software, namely Sketch Engine. The study focuses on the frequency, structural and functional characteristics of lexical bundles using a mixed-method research design. The English corpus comprises 1,351,048 words derived from 124 RA, while the Indonesian corpus consists of 637,910 words collected from 124 RA. We found that three-word lexical bundles are more prevalent than four-word lexical bundles in both corpora. Based on the structural forms, prepositional-based bundles are the most frequent form in English RA, while noun-based bundles are the most common form in Indonesian RA. There were no participant-oriented bundles found in the Indonesian RA corpus in terms of functional classification, whereas the English RA corpus involved more varied functional categories of lexical bundles. The findings provide an understanding of phraseological combinations in English and Indonesian scientific writing, characterizing disciplinary discourse as well as native and non-native English speakers' rhetorical style, and have pedagogical implications for EAP practitioners.

Keywords: corpus, frequency, lexical bundles, phraseology, research article.

ABSTRAK

Rendahnya tingkat publikasi para peneliti Indonesia di jurnal internasional bereputasi, terutama dalam bidang seni dan humaniora, mungkin disebabkan oleh gaya retorika yang berbeda dalam artikel ilmiah berbahasa Indonesia dan Inggris. Artikel ini mengkaji simpul leksikal dalam korpus artikel ilmiah berbahasa Inggris dan Indonesia tentang sastra dan linguistik, dengan menggunakan rancangan metode gabungan. Kajian berfokus pada pembahasan frekuensi penggunaan, struktur, dan fungsi sampul leksikal dalam korpus. Korpus bahasa Inggris terdiri atas 1.351.048 kata yang diperoleh dari 124 artikel di jurnal internasional, sedangkan korpus bahasa Indonesia terdiri atas 637.910 kata yang dikumpulkan dari 124 artikel di jurnal nasional. Berdasarkan hasil analisis, kami menemukan bahwa pada kedua korpus simpul leksikal yang terdiri atas tiga kata lebih banyak daripada yang empat kata. Berdasarkan strukturnya, korpus artikel berbahasa Inggris didominasi oleh bentuk simpul berbasis preposisi, sedangkan korpus artikel berbahasa Indonesia memiliki lebih banyak simpul berbasis nomina. Dari fungsinya, simpul yang berorientasi partisipan tidak ditemukan dalam bahasa Indonesia, sedangkan dalam bahasa Inggris simpul leksikal memiliki fungsi yang lebih beragam. Hasil penelitian ini memberikan kontribusi pada pemahaman tentang kombinasi fraseologi dalam penulisan ilmiah berbahasa Inggris dan bahasa Indonesia, yang mencirikan wacana disipliner dan juga gaya retoris penutur jati dan penutur nonjati bahasa Inggris, serta memiliki implikasi pedagogis untuk para praktisi di bidang bahasa Inggris untuk tujuan akademik.

Kata kunci: *frekuensi, korpus, simpul leksikal, artikel ilmiah.*

INTRODUCTION

Ministry of Research, Technology, and Higher Education of the Republic of Indonesia reported in 2016 that the lowest number of academic publications is from the fields of arts and humanities (0.91%), while the highest is from the fields of science, technology, health, and medicine (15,14%) (Arsyad, Purwo, Sukamto, & Adnan, 2019). The report suggests that researchers in arts and humanities have published the fewest research articles (RA) in reputable international journals compared to researchers in the other fields. One of the possible reasons hindering them from publishing scientific articles is their difficulties in writing accurate and effective expository texts in English that are different from those in Indonesian.

The reason might be a cliché, but it is undeniable that a significant number of non-native researchers from all over the world are facing the fact that English plays a central role in disseminating academic knowledge. Consequently, they have been struggling with a lack of proficiency in English and unfamiliarity with the standard rhetorical style expected in English journal articles. The difficulties are challenged by non-native scientists who have encouraged scholars to conduct many studies on the elements that create well-written academic prose. Some insightful studies used corpora, large bodies of machine-readable text, to investigate the linguistic forms and discourse structures within particular texts or genres.

Corpus-based language studies have encouraged a paradigm shift in learning English as a foreign language, specifically for adult learners. From the traditional perspective, words are thought of as the basic building blocks of language learning and processing. Therefore, some of the research recommended vocabulary and lexical approach as the ground for learning a foreign language (Wilkins, 1972; Harmer, 1991; & Lewis, 1993). However, recent theories and empirical evidence show that multi-word sequences are the integral building blocks for language. Additionally, the predominance of multi-word sequences in a discourse shows that meaning creation and understanding largely depend on stocks of the multi-word sequences in language users' lexicon (Sinclair, 1991 and Hong & Hua, 2018). For this reason, studies on multi-word expressions and lexicon in a variety of registers have been flourishing in recent years.

Multi-word sequences have significantly been studied under many rubrics, for example, phraseological sequences, formulaic language, chunks, clusters, multi-word units, recurrent sequences, recurrent word combinations, lexical phrases, formulas, routines, fixed expressions, prefabricated patterns (prefabs), phrasicon, n-grams, and lexical bundles (Biber, Conrad & Cortes, 2004; Hong & Hua, 2018; & Hernandéz, 2013). According to Biber, Conrad & Cortes (2004) and Biber, Johansson, Leech, Conrad, & Finegan (1999), lexical bundles are multi-word units that occur with a high frequency in a register. They specifically define that lexical bundles are "bundles of words that show a statistical tendency to co-occur" (Biber, Johansson, Leech, Conrad, & Finegan, 1999: 989).

Salazar (2014) explains that the main feature of lexical bundles is that they have an empirical basis due to the method of determination which primarily depends on frequency criteria. Therefore, she defines lexical bundles as "frequently occurring lexical sequences automatically extracted from a given corpus using a computer program" (Salazar, 2014: 13). Lexical bundles are regarded as the fundamental part of a discourse that plays a significant role in creating fluency and achieving the natural use of language, either in speech or writing (Kashiha, 2015). As a result, many studies have investigated the relations between lexical bundles and language proficiency.

Millar (in Allen, 2011) argued that the knowledge and use of various lexical bundles could help language learners attain naturalness in language use. On the contrary, the misapplication of lexical bundles is shown to be a potential cause of communication problems. Besides, some studies showed that language learners with a higher frequency of lexical bundles demonstrated higher language proficiency (Novita & Kwary, 2018). The knowledge about the high frequent lexical bundles and the patterns

of use in scientific writing of a specific discipline are essential for non-native writers because they are highly expected to produce brief and accurate explanatory texts to communicate their thoughts and research findings to a worldwide scientific audience.

Due to their importance in language learning for academic purposes, there have been many studies on lexical bundles used by the first language (L1) and second language (L2) writers in academic genres. For example, Chen and Baker (2010) conducted research on frequently-used lexical bundles in L1 and L2 academic writing and argued that the frequency-driven lexical bundles found in native expert writing could greatly assist learner writers in achieving a more native-like style of academic writing. Salazar (2014) compared the use of lexical bundles in a corpus of biomedical RA written by native Spanish-speaking scientists with a corpus of health science RA written by English native speakers. Kashiha (2015) examined lexical bundles in two different corpora of RA conclusion sections of native and Iranian non-native English. Pan, Reppen, & Biber (2016) studied lexical bundles in the context of the structural and functional types used by L1 English and L1 Chinese professional writing in Telecommunications journals.

Nevertheless, no research to date has compared the lexical bundles of RA from different languages. The current study addresses to fill the gap by investigating the frequency of use, structural and functional characteristics of lexical bundles in English and Indonesian RA. The study aims to compare lexical bundles in the same genre and discipline, which are literature and linguistics, but written in different languages to reveal fundamental similarities and differences in terms of frequency and patterns of multi-word expressions. In this context, the present study focuses on the formulaic language in published RA of Indonesian and English instead of language proficiency. Hence, this study can demonstrate the norm of language use in scientific writing of Indonesian and English as well as to gain an understanding of the linguistic aspects hindering Indonesian researchers from publishing RA in reputable international journals.

METHOD

Data for this study are two corpora of written texts comprising Indonesian and English RA from literature and linguistics, which are open access articles. The Indonesian RA corpus was built from Indonesian national journals indexed in Science and Technology Index (SINTA) from the Ministry of Research and Technology of the Republic of Indonesia from SINTA 1 to SINTA 3. The corpus consists of 124 published RA in the leading journals of each category.

On the other hand, the English RA corpus was collected from international journals indexed in Scopus with the category of Q1 and Q2, comprising 124 published articles. From the same number of articles we collected, the size of the corpora is different. As shown in Table 1, the English RA corpus is two times bigger than the Indonesian RA corpus. The corpus size suggests that the number of words of articles published by Indonesia's reputable journals is generally smaller than those published by reputed international journals. Indonesia's journal publishers may consider this to achieve a more standard quality of international journals.

TABLE I CORPORA WORD COUNTS

No	Corpus	Number of Articles	Number of Tokens	Number of Types
1	Indonesian RA	124	637,910	47,938
2	English RA	124	1,351,04	59,771

We extracted lexical bundles of the corpus data using corpus software, namely Sketch Engine (Kilgarriff et al., 2014). The software was used to generate the most frequent lexical bundles in both corpora ranging from 3-word bundles to 5-word bundles for frequency analysis. However, we focused on 4-word bundles for structural and functional analyses. The determination is based on the research conducted by Hyland (2008), stating that 4-word and 5-word bundles provide a more precise range of structures and functions than 3-word bundles. In selecting the lexical bundles with a high frequency, we also set a minimum frequency of 20.

The data analyses consist of several steps. First, we compared the pattern of the top 50 most frequent lexical bundles in the corpora of English and Indonesian published RA in terms of frequency. Second, we chose the 4-word bundles to 5-word bundles in the top 50 most frequent lexical bundles and categorized them based on the structure or grammatical types and the function or their meaning in the texts. The structural classification of lexical bundles follows the taxonomy developed by Biber, Johansson, Leech, Conrad, & Finegan (1999), consisting of noun-based, prepositional-based, and verb-based bundles.

On the other hand, the functional classification of lexical bundles refers to the category initially created by Biber (2006) and Biber, Conrad, & Cortes. (2004) and then modified by Hyland (2008 & 2012), which consists of research-oriented, text-oriented, and participant-oriented. The research-oriented bundles "help writers to structure their activities and experiences of the real world (Hyland, 2012: 150), which subcategories are location, procedure, quantification, description, and topic. The text-oriented bundles involve "the organization of the text and its meaning as a message or argument" (Hyland, 2012: 150). The subcategories of this function are transition, resultative, structuring, and framing signals. The participant-oriented bundles pay particular attention to the reader or writer of the text, consisting of stance and engagement features (Hyland, 2012: 150). Based on these analysis results, we compared and interpreted the pattern of lexical bundles in the corpora of English and Indonesian published RA.

RESULTS AND DISCUSSION

In the present study, the description of lexical bundles in the corpora greatly depends on frequency criteria. It follows the way Biber, Johansson, Leech, Conrad, & Finegan (1999) investigated lexical bundles, which is exclusively grounded in the frequency. The analysis is based on the idea that frequency provides strong evidence of the characteristic combinations and primary meaning of words in specific contexts (Hunston, 2006). This approach certainly helps us analyze and compare lexical bundles' structure and function in two different languages of the same genre.

TABLE II TOP 50 MOST FREQUENT LEXICAL BUNDLES IN ENGLISH AND INDONESIAN PUBLISHED RA

No	Item	Normalized Freq.	No	Item	Normalized Freq.
1	as well as	10,846	1	*dalam bahasa Indonesia*	18,585
2	the use of	8,798	2	*dalam penelitian ini*	17,085
3	one of the	7,192	3	*oleh karena itu*	16,694
4	in terms of	6,866	4	*penelitian ini adalah*	11,216
5	in order to	6,866	5	*di bawah ini*	10,303
6	the fact that	5,912	6	*dalam hal ini*	10,108
7	in which the	4,562	7	*yang ada di*	9,977
8	the end of	4,329	8	*yang dilakukan oleh*	9,325
9	on the other	3,840	9	yang digunakan dalam	8,086

10	a number of	3,840	10	*dengan kata lain*	7,825
11	the number of	3,747	11	*merupakan salah satu*	7,630
12	in other words	3,747	12	*yang berkaitan dengan*	7,173
13	there is a	3,677	13	*yang terdapat dalam*	5,999
14	part of the	3,608	14	*makian dalam bahasa*	5,934
15	the United States	3,584	15	*makian dalam bahasa Indonesia*	5,869
16	of the novel	3,584	16	*yang berasal dari*	5,739
17	the same time	3,491	17	*dalam penelitian ini adalah*	5,412
18	it is not	3,468	18	*oleh sebab itu*	5,086
19	at the same	3,445	19	bahasa Indonesia yang	4,956
20	the present study	3,305	20	*ini menunjukkan bahwa*	4,826
21	in relation to	3,305	21	*laki-laki dan perempuan*	4,695
22	at the same time	3,305	22	*anak disabilitas tunarungu*	4,695
23	the case of	3,189	23	*yang digunakan oleh*	4,500
24	the context of	3,119	24	*yang berhubungan dengan*	4,500
25	such as the	3,119	25	*makian dengan referensi*	4,500
26	end of the	3,049	26	*bahasa Minangkabau Bukittinggi*	4,500
27	of the world	3,026	27	*Nyi Roro Kidul*	4,434
28	to be a	3,003	28	*klitika pronominal pemarkah*	4,434
29	can not be	2,979	29	yang digunakan untuk	4,369
30	in the first	2,956	30	*dapat disimpulkan bahwa*	4,304
31	the role of	2,863	31	*dan berbau harum*	4,304
32	in the context	2,746	32	*yang berada di*	4,108
33	use of the	2,677	33	*pronomina pemarkah kasus*	4,108
34	the other hand	2,653	34	*klitika pronomina pemarkah kasus*	4,108
35	the importance of	2,630	35	*dalam bahasa Inggris*	4,043
36	the end of the	2,630	36	*adalah salah satu*	4,043
37	on the other hand	2,630	37	*yang terkait dengan*	3,978
38	some of the	2,560	38	*di samping itu*	3,978
39	of world literature	2,560	39	sebagai bagian dari	3,913
40	in the context of	2,537	40	*digunakan dalam penelitian*	3,913
41	in the case	2,537	41	*menjadi salah satu*	3,847
42	the relationship between	2,514	42	*dapat dikatakan bahwa*	3,717
43	the waste land	2,490	43	*yang digunakan dalam penelitian*	3,652
44	in this study	2,444	44	*sebagai salah satu*	3,652
45	a variety of	2,444	45	*dapat dilihat pada*	3,652
46	a kind of	2,444	46	*yang ada dalam*	3,521
47	that it is	2,421	47	*digunakan dalam penelitian ini*	3,521
48	understanding of the	2,397	48	*dalam bahasa Jawa*	3,456
49	in the case of	2,374	49	yang terjadi di	3,391
50	the form of	2,351	50	*ibu rumah tangga*	3,326

Based on the method described in the previous section, the focus of the analysis is the top 50 most frequent lexical bundles in English and Indonesian corpora of published RA. As shown in Table II, the lexical bundles in high frequency consist of 3-word bundles and 4-word bundles, and the lists are mainly composed of three-word strings. In other words, the 3-word bundles are more productive not only in English but also in the Indonesian RA corpus. However, the English corpus has slightly more 3-word lexical bundles than the Indonesian corpus. It can be seen from the number of 4-word bundles in both of the corpora. The English corpus has only two 4-word bundles, which are *on the other hand* and *in the context of*, while the Indonesian corpus has five 4-word bundles, which are *makian dalam bahasa Indonesian*, *dalam penelitian ini adalah*, *klitika pronomina pemarkah kasus*, *yang digunakan dalam penelitian*, dan *digunakan dalam penelitian ini.*

As expected, the result of frequency analysis is in line with what was stated by Hyland (2012), who studied lexical bundles in academic discourse. According to him, 3-word bundles are exceedingly prevalent, but they are often less interesting to investigate further. In this context, the most important thing to note is that the pattern of lexical bundles in the corpora of English and Indonesian published RA is similar in terms of the frequency of use.

After comparing the lexical bundles in the English RA corpus with the Indonesian RA corpus from the aspect of frequency, it will also be much more insightful if we investigate them from the structural forms. As stated in the method section, the structural classification is based on the taxonomy developed by Biber, Johansson, Leech, Conrad, & Finegan (1999), who divided the structural forms into three broad structural categories, namely noun-based, prepositional-based, and verb-based bundles. NP-based bundles comprise any nouns with post-modifier fragments, PP-based bundles include any word combinations initiated by preposition followed by noun phrase fragments, and verb-based bundles refer to a string of words with verb components. The structural forms of lexical bundles in the corpus of English language RA are shown below in Table III, while in the corpus of Indonesian language RA is presented in Table 4.

TABLE III THE STRUCTURAL FORMS OF LEXICAL BUNDLES IN ENGLISH PUBLISHED RA

	Structural forms	types	% of types		Lexical bundles
Noun-based	Noun phrase with of- phrase fragment	10	20%	1	the end of
				2	the rest of the
				3	the use of the
				4	one of the most
				5	the beginning of the
				6	a wide range of
				7	the total number of
				8	the case of the
				9	the nature of the
				10	the context of the
	Noun phrase with other post-modifier fragment	4	8%	11	the extent to which
				12	the ways in which
				13	the fact that the
				14	the way in which
	Total	14	28%		

Prepositional-based	Prepositional-based with embedded -of phrase	18	36%	15	in the context of
				16	in the case of
				17	at the end of
				18	in the form of
				19	on the basis of
				20	at the end of the
				21	in terms of the
				22	in the use of
				23	as a result of
				24	on the part of
				25	in the face of
				26	at the university of
				27	over the course of
				28	at the beginning of
				29	at the heart of
				30	of the waste land
				31	of look to the
				32	of the singular marker
	Prepositional-based with other post-modifier fragment	4	8%	33	to the fact that
				34	in a way that
				35	by the fact that
				36	in the sense that
	Other prepositional phrase segments	9	18%	37	at the same time
				38	on the other hand
				39	on the one hand
				40	in the United States
				41	in the present study
				42	in relation to the
				43	in the same way
				44	with respect to the
				45	with regard to the
	Total	31	62%		
Verb-based	Verb phrase with active verb	1	2%	46	looks to the subject
	Be+noun phrase	1	2%	47	is one of the
	Passive verb	1	2%	48	can be seen in
	Verb/adjective+to	1	2%	49	It is important to
	Adverbial clause	1	2%	50	as well as the
	Total	5	10%		

The data in Table III reveal that lexical bundles in English RA are primarily in the form of prepositional-based and noun-based bundles. The prepositional-based bundles are sixty-two percent, and the noun-based bundles are twenty-eight percent, making a total of ninety percent. The lowest number of structural forms is verb-based bundles, which are only ten percent. The results are similar to the research findings shown by Hyland (2008), who analyzed doctoral dissertations across four disciplines (electrical engineering, business studies, applied linguistics, and microbiology), Jalali, Moini, & Arani (2015), who studied medical research articles, Pan, Reppen, & Biber (2016), who investigated research articles in telecommunications research journals, and other previous research conducted by Dontcheva-Navratilova (2012), Bal (2010) and Liu (2008) who examined a variety of academic registers. They discovered that the most common lexical bundles are prepositional-based and noun-based. The results of the present study are slightly different from those found by Kwary, Ratri, & Artha (2017) and Qin (2014), who analyzed lexical bundles in journal articles across four disciplines (life sciences, health sciences, physical sciences, and social sciences) and applied linguistics respectively. They found that prepositional-based is the most frequent bundles, but the verb-based is the second frequent one instead of noun-based bundles.

It is also important to note that the prepositional-based bundles are predominantly prepositional phrases with embedded phrase fragments, i.e., 18 out of 31 types, such as *in the context of*, *in the case of*, *on the basis of*, and *in terms of the*. These structural forms typically relate to the text structure and its meaning, especially to establish arguments by describing limiting conditions. On the other hand, the noun-based bundles are mainly noun phrases with of phrase fragments, i.e., 10 out of 14 types, for example, *the end of*, *the rest of the*, *the use of*, and *one of the most.* These forms function to help writers organize their activities and experiences of the real world by indicating time/ place, quantity, and procedure.

TABLE IV THE STRUCTURAL FORMS OF LEXICAL BUNDLES IN INDONESIAN PUBLISHED RA

Structural forms		types	% of types		Lexical bundles
Noun-based	Noun phrase segments	4	8%	1	klitika pronomina pemarkah kasus
				2	*kongres bahasa Indonesia I*
				3	*wayang orang Ngesti Pandowo*
				4	*tari Bedhaya Bedhah Madiun*
	Noun phrase with post- modifier fragment	29	58%	5	*ragam tutur yang lebih*
				6	*satu dengan yang lain*
				7	*panas dan berbau harum*
				8	*tanah panas dan berbau harum*
				9	*tanah panas dan berbau*
				10	*tanah hangat dan berbau*
				11	*tutur yang lebih kasual*
				12	*tanah hangat dan berbau harum*
				13	*ragam tutur yang lebih kasual*
				14	*anak disabilitas tunarungu usia*
				15	*Jamee dan bahasa Minangkabau Bukittinggi*

				16	*Jamee dan bahasa Minangkabau*
				17	*bahasa Jamee dan bahasa Minangkabau*
				18	*bahasa Jamee dan bahasa*
				19	*makian dalam bahasa Indonesia*
				20	*data dalam penelitian ini*
				21	*makian dengan referensi binatang*
				22	*penggunaan makian dalam bahasa*
				23	*penggunaan makian dalam bahasa Indonesia*
				24	*nama-nama geng sekolah di*
				25	*geng sekolah di Yogyakarta*
				26	*nama-nama geng sekolah di Yogyakarta*
				27	*penggunaan ragam tutur yang*
				28	*kata tabu yang berhubungan*
				29	*ungkapan yang mengandung sikap seksis*
				30	*ungkapan yang mengandung sikap*
				31	*tabu yang berhubungan dengan*
				32	*kata tabu yang berhubungan dengan*
				33	*hangat dan berbau harum*
	Total	33	66%		
Prepositional-based	Prepositional-phrase segments	6	12%	34	*oleh klitika pronomina pemarkah kasus*
				35	*oleh klitika pronomina pemarkah*
				36	*dalam bahasa Indonesia yang*
				37	*ke dalam bahasa Indonesia*
				38	*yakni penggunaan ragam tutur*
				39	*dalam penelitian ini adalah*
	Total	6	12%		
Verb-based	Passive verb	5	10%	40	*yang digunakan dalam penelitian*
				41	*digunakan dalam penelitian ini*
				42	*yang digunakan dalam penelitian ini*
				43	*digunakan dalam penelitian ini adalah*
				44	*dimarkahi oleh klitika pronomina*

	Active verb	5	10%	45	*menggunakan makian dengan referensi*
				46	*yang mengandung sikap seksis*
				47	*abstrak penelitian ini bertujuan*
				48	*this study aims to*
				49	*hasil penelitian menunjukkan bahwa*
	Total	10	20%		
Other		1	2%	50	*dan bahasa Minangkabau Bukittingi*

On the other hand, the Indonesian RA corpus mostly comprises noun-based and verb-based bundles, as shown in Table IV. Sixty-six percent of clusters are noun-based, whereas twenty percent are verb-based, making a total of eighty-six percent. The lowest number of structural types is prepositional-based, which is twelve percent, and other categories, which are two percent. The results suggest that the patterns of lexical bundles in Indonesian RA are different from those found in English RA in terms of structural classification. As discussed before, English research articles have more prepositional-based (62%), while Indonesian research articles have more noun-based (66%). The noun-based bundles are pretty common in English RA, but the distribution differs from the Indonesian RA. The noun-based bundles in English RA (14 types) are less than half of those found in the Indonesian RA (33 types). These results, in general, are also different from the findings shown in the research conducted by Hyland (2008), Qin (2014), Jalali, Moini, & Arani (2015), Pan, Reppen, & Biber (2016), and Kwary, Ratri, & Artha (2017) who found that the most frequent bundles are prepositional-based. Thus, the differences are possibly caused by the difference in terms of language rather than the fields of study.

If we examine further, the noun-based bundles in Indonesian RA are mostly noun phrases with post-modifier fragments, for example, *ragam tutur yang lebih*, *panas dan berbau harum*, *tanah panas dan berbau harum*, dan *tutur yang lebih kasual*. Writers typically use these to structure their activities and experiences, mainly related research topics. From this function, it can also be seen that the noun-based bundles in Indonesian RA and English RA are different in subcategories of the structural forms as well as the function.

TABLE V FUNCTIONAL CLASSIFICATION OF LEXICAL BUNDLES IN ENGLISH AND INDONESIAN PUBLISHED RA

Function	English		Indonesian	
	Types	% of Types	Types	% of Types
Research-oriented bundles	29	58%	46	92%
Location	7		-	
Procedure	2		4	
Quantification	6		-	
Description	8		-	
Topic	6		42	

Text-oriented bundles	19	38%	4	8%
Transition signals	4		-	
Resultative signals	1		1	
Structuring signals	1		3	
Framing signals	13		-	
Participant-oriented bundles	2	4%	-	-
Stance features	1		-	
Engagement features	1		-	

As stated in the method section, the functional classification of lexical bundles in this current study refers to the classification proposed by Hyland (2008 & 2012). The results show that the function of lexical bundles in English and Indonesian RA shares some similarities and differences. The research-oriented bundles are found to be the most frequent category in both English and Indonesian RA, but the distribution differs. In the English RA corpus, the functional type of research-oriented bundles is fifty-eight percent, while in the Indonesian RA corpus, it is ninety-two percent. It suggests this type of function much more dominates the Indonesian RA. The rest, which is eight percent, is text-oriented bundles, while participant-oriented bundles are not found. As shown in Table V, the research-oriented bundles were mainly used to impart the research topics. Many of these bundles specified the subject of the research. They were realized by noun phrase structure, such as *makian dalam bahasa Indonesia, klitika pronomina pemarkah kasus, Kongres Bahasa Indonesia I, wayang orang Ngesti Pandowo, tari bedhaya bedhah Madiun, ragam tutur yang lebih, geng sekolah di Yogyakarta,* and *makian dengan referensi binatang*.

In contrast, the word combinations functioning as research-oriented bundles in English RA corpus are lower, i.e., 58% and their types are not dominated by topic; they are more varied instead. The bundles are mainly used to describe objects, relation, and degree, for example, *in the form of, the ways in which, in relation to the,* and *the extent to which*. Many of them also contribute to the description of location, such as *the end of the, at the end of, the beginning of the, at the beginning of the,* and *at the heart of*. Meanwhile, lexical bundles functioning to explain procedure are the least, e.g., *in the use of* and *the use of.*

Furthermore, the number of text-oriented bundles in the English RA corpus is relatively high, i.e., 38%. They primarily function to frame arguments by showing limitation, describing connection, and specifying cases, such as *in the context of, in the case of, on the basis of, in terms of, the fact that, in the sense that, with respect to the, with regard to the,* and *the case of the.* As can be seen, these bundles are realized mainly by preposition with embedded -of phrase structure. The other kind of bundles in the text-oriented category that is found quite many is transition signals, e.g., *as well as the, on the other hand,* and *in the one hand*. These are mainly used to link arguments in a logical order by introducing additional information and contrasting a point of view. The category of resultative signals is also found in the data, e.g., *as a result of.* According to Hyland (2008), transition words, particularly the resultative markers, for instance, *as a result of*, is a crucial function in rhetorical presentation of research because they signal the main conclusions from the research and emphasize the inferences the writers want readers to draw from the discussion.

The most notable difference between the English RA corpus and Indonesian RA corpus is in terms of participant-oriented bundles. As mentioned before, none of the participant-oriented bundles is found in the Indonesian RA corpus. However, in the English RA corpus, we

found stance features, i.e., *it is important to*, and engagement features, i.e., *can be seen in*, in the participant-oriented category. Hyland (2008) stated that stance features relate to the ways writers explicitly intervene into the discourse to communicate epistemic and evaluative judgment, evaluations, and degrees of commitment to what they tell, while engagement features concern the ways the writers address readers as participants in the unfolding discourse. In line with that statement, in the English RA corpus, the bundle *it is important* is mainly used to convey the writers' evaluation of what they believe to be essential to note and consider. Meanwhile, the use of the bundle *can be seen in* demonstrates the way the writers want readers to recognize. Thus, the participant-oriented bundles used in the corpus of English RA are a part of the dialogic element of research writing to direct the readers to some understanding, which is not found in the Indonesian RA corpus.

CONCLUSION

The main objective of the current study is to explore the patterns of lexical bundles in the corpora of English and Indonesian RA, built from published scientific articles in the fields of literature and linguistics, by using a corpus tool, namely the Sketch Engine. By analyzing the frequency, structural forms, and functional classification, lexical bundles in English RA and Indonesian RA corpora show some similarities and differences. Based on the top 50 most frequent lexical bundles, the results show that the number of three-word bundles is higher than four-word bundles in both English and Indonesian RA corpora. The results strengthen findings revealed by Hyland (2012) that three-word bundles are the most common bundle found in English academic discourse and proven that this typical lexical bundle occurs not only in English but also in Indonesian academic discourse.

The most notable differences found between English RA and Indonesian RA are in the case of structural forms and the distribution of functional categories of four-word bundles. While the English RA corpus is dominated by prepositional-based bundles (62%), the Indonesian RA corpus is mostly noun-based bundles (66%). Furthermore, the second most common types of structural forms in both corpora are different, i.e., noun-based bundles in English RA corpus (28%) and verb-based bundles in Indonesian RA corpus (20%). The findings suggest that in terms of structure, there are differences between the way writers write articles in English and Indonesian.

The other differences between English RA and Indonesian RA corpora can be seen from the functional classification. Although the research-oriented bundles are the most common type found in both corpora, the distribution of the type and its subcategories differs. Indonesian RA corpus has a more significant number of research-oriented bundles (92%) than the English RA corpus (58). Besides, the research-oriented bundles in English RA are more varied, including all the subcategories, i.e., location, procedure, quantification, topic, and description. In contrast, in the Indonesian RA corpus, the research-oriented bundles are predominantly topic (92%). Unlike the English RA corpus, the Indonesian RA corpus has no participant-oriented bundles. It indicates that the writers in Indonesian RA tend not to show a dialogic aspect with their readers.

The study demonstrates how technology, in this case, the corpus tool Sketch Engine, has greatly facilitated researchers to identify the phraseological pattern in a large sample collection of language use and indicates writers of native and non-native English use different rhetorical styles. However, the findings need to be considered with some caution because we analyzed based on relatively limited kinds and the number of data and have not deeply discussed the data in terms of rhetorical style in the related discipline as well as the discourse style in the related languages. In spite of that, the results have clear pedagogic implications for English for Academic Purposes practitioners, especially those who teach EAP for Indonesian EFL. The findings can be used as the source of learning materials about the phraseological forms in English scientific articles as well as the norm of language in academic English in general.

REFERENCES

Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education*, *1*, 105-127.

Ang, L. H., & Tan, K. H. (2018). Specificity in English for Academic Purposes (EAP): A Corpus analysis of lexical bundles in academic writing. *3L: Language, Linguistics, Literature®*, *24*(2).

Arsyad, S., Purwo, B. K., Sukamto, K. E., & Adnan, Z. (2019). Factors hindering Indonesian lecturers from publishing articles in reputable international journals. *Journal on English as a Foreign Language*, *9*(1), 42-70.

Bal, B. (2010). *Analysis of Four-word Lexical Bundles in Published Research Articles Written by Turkish Scholars*. Georgia State University.

Biber, D. (2006). *University language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

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

Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman Grammar of Spoken and Written English.* Pearson Education, Ltd.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (2000). *Longman grammar of spoken and written English*. Longman.

Dontcheva-Navratilova, O. (2012). Lexical bundles in academic texts by non-native speakers. *Brno Studies in English,* 38-2, 37-58.

Harmer, J. (1991). *Teaching vocabulary: The practice of English language teaching* (2nd ed.). Longman.

Hernández, P. S. (2013). Lexical bundles in three oral corpora of university students. *Nordic Journal of English Studies*, *12*(1), 187-209.

Hunston, S. (2006). *Corpora in Applied Linguistics*. Cambridge University Press.

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

Hyland, K. (2012). Bundles in academic discourse. *Annual review of applied linguistics*, *32*, 150-169.

Jalali, Z. S., Moini, M. R., & Arani, M. A. (2014). Structural and functional analysis of lexical bundles in medical research articles: A corpus-based study. *International Journal of Information Science and Management (IJISM)*, *13*(1).

Kashiha, H. (2015). Recurrent formulas and moves in writing research article conclusions among native and non-native writers. *3L: Language, Linguistics, Literature®*, *21*(1). 47-59.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*(1), 7-36.

Kwary, D. A., Ratri, D., & Artha, A. F. (2017). Lexical bundles in journal articles across academic disciplines. *Indonesian Journal of Applied Linguistics*, *7*(1), 131-140.

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward.* Language Teaching Publications.

Liu, D. (2012). The most frequently-used multi-word constructions in academic written English: A multi-corpus study. *English for Specific Purposes*, *31*(1), 25-35.

Novita, H., & Kwary, D. A. (2018). Comparing the use of lexical bundles in Indonesian-English translation by student translators and professional translators. *Translation & Interpreting, The*, *10*(1), 53-74.

Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes*, *21*, 60-71.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching* (Vol. 65). John Benjamins Publishing Company.

Sinclair, J., & Sinclair, L. (1991). *Corpus, concordance, collocation*. Oxford University Press, USA

Wilkins, D. A. (1972). *Linguistics in Language Teaching*. Arnold.