JOMANTARA
INDONESIAN JOURNAL OF ART AND CULTURE

FISS. Vol. 4, No.1, January 2024

An Analysis of Lexical Bundles in Research Articles: Examples from Natural and Social Sciences

Annur Karima Zulyanputri[1]

[1] Universitas Padjadjaran
Sumedang
annur19001@mail.unpad.ac.id

Abstract

This present research investigates the use of lexical bundles in research articles across different disciplinary areas: natural science and social science. The results showed that there are some differences in the use of lexical bundles between disciplines in the term of frequency, structures, and functions. We found that the frequency of lexical bundles can vary between natural science and social science RA. For example, the lexical bundle "*in the learning process*" is more common in natural science, while the lexical bundle "*on the other hand*" is more common in social science. Based on the structural forms, the most common structural form of lexical bundles in natural science RA is verb-based. While the most common structural form of lexical bundles in social science RA is prepositional-based. The contradiction suggests that in terms of structure, there are differences between natural and social science RA writing styles, how the authors establish arguments, and also present results. Meanwhile from the functional classifications, we can conclude that both types of RA have the same order in terms of the frequency of the functional use, with research-oriented bundles being the most frequent. The findings can be used to improve the readability of research articles in one discipline, as well as to help researchers learn the conventions of writing in one discipline.

Keywords: lexical bundles, research articles, frequency, natural science, social science.

1. Introduction

Writing scientific research in English for most ESL researchers can be a challenging process, as they have to put out more effort into finding the right words to construct their sentences than native English speakers. Meanwhile, scientific writing plays a key role in the academic context because scholarly publications contribute to career promotion and reputation (Thanh Tuyen et al., 2016). For better readability research, developing writing skills is necessary for ESL researchers. This could be maintained by considering a few things that inhibit writing development such as limited vocabulary and poor understanding of grammar.

Scientific research is essentially a type of communication for students, lecturers, researchers, or scientists to convey or publish their thoughts,

analysis, and findings. It refers to a particular style of expression used by the authors to define the boundaries of their disciplines and areas of expertise. Different disciplines also have different writing styles and structures. For example, some disciplines such as humanities expect longer paragraphs, which include a topic sentence to show how an argument is structured. Meanwhile, other disciplines such as sains need to incorporate a lot of numbers and units into the writing. Therefore, reading an academic paper could also help the reader define the author's discipline.

To produce natural and meaningful texts, formulaic expressions are required. In linguistics, formulaic expressions are also known as prefabricated language. They are a type of pragmatic language, which means that they are used to achieve a specific communicative purpose. Formulaic expressions often used in spoken language, but it can also be found in written language. Some examples of formulaic expressions include: "*thank you*", "*how are you*", and "*see you later*". Baker & Chen (2010) specified formulaic expression as formulaic sequence/language, it is an umbrella term often used to refer to various types of multi-word units. The term formulaic language is then used by Wood (2019) to define multiword language phenomena which holistically represent a single meaning or function. By looking at some research, it is acknowledged that researchers develop the thought of formulaic expression in different ways, though the phenomenon is the same; formulaic expression covers various lexical units including idioms, proverbs, collocations, lexical bundles, and other conventional and multi-words units.

This present research is then written to investigate lexical bundles as one of the categories of formulaic expression since it is naturally used in both spoken and written language, besides collocation. The lexical bundle itself, defined as an expression of a sequence of three or more words that frequently recurs in natural discourse

regardless of their idiomaticity and structural status (Biber et al., 1999 in Budiwiyanto & Suhardijanto, 2020). This object has been widely discussed in previous researches. Along with its usage in everyday language, studies on lexical bundles also can be conducted in spoken registers such as conversations, group discussions, and lectures, and written registers such as textbooks, students' projects, and research articles. However, this present research focuses on analysing written registers, particularly research articles across different disciplinary areas. Moreover, Salazar (2014) added that it is important for a second or foreign-language writer to know the most frequent combinations used in specific registers, genres, and disciplines. As Indonesian people are not English native speakers, various factors affect their understanding when producing group of words in both spoken and written language. The possible outcome that occurs is that this condition produces various styles of how Indonesian people, especially researchers produce group of words/multi-word units in their scientific writing.

There are numerous previous researches regarding cross-disciplinary investigation of lexical bundles. Kashiha & Heng (2013) offered research of lexical bundles that used in 24 academic lectures of hard and soft sciences, taken from the British Academic Spoken English (BASE). Research by Kwary et al., (2017) focused on the use of lexical bundles in journal articles of four academic disciplines stated in Scopus: health sciences, life sciences, physical sciences, and social sciences. Another research of lexical bundles focused on certain discipline conducted by Budiwiyanto & Suhardijanto (2020), they identified Indonesian lexical bundles of six disciplines. To fill the gap, this present research will be focused on analysis of lexical bundles in research articles that written by Indonesian authors from four disciplinary areas: Medical Science, Mathematics & Natural Science, Linguistics, and Humanities & Social Science.

2. Method

The corpus of the present research is gathered from research articles written by Indonesian researchers. There are four corpora consisting of research articles from four different disciplinary areas, namely: (1) Linguistics, (2) Humanities & Social Science, (3) Medical Science, and (4) Mathematics & Natural Science. These disciplinary areas represent two branches of science (i.e., natural science and social science) and the distribution is based on the data of *lldikti12.ristekdikti.go.id* (Indonesian Ministry of Research and Technology). With a total of 200 research articles, each discipline has 50 research articles that are indexed in Science and Technology Index (SINTA), ranging from SINTA 1 to SINTA 2. The corpus comprises 857.706 words after going through the data-cleaning process. Each article that was downloaded in *pdf* format converted to *docx*, aimed to clear up all unintended information including journal volume description, the author's name and affiliation, and references. The data is henceforth converted into *txt* or plain text format since it is compatible with Antconc (Anthony, 2004). Below is the table of tokens (number of words) in the corpus.

Table 1. Corpus Size

No	Corpus	Number of Articles	Number of Words
1	Linguistics RA	50	241.544
2	Humanities & Social Science RA	50	307.273
3	Medical Science RA	50	135.825
4	Mathematics & Natural Science RA	50	173.064
Total		200	857.706

To build the corpus, several criteria are set: (1) the articles must be open articles, which means that the articles could be freely accessed and downloaded, (2) the author is an Indonesian researcher who is associated with universities or institutions in Indonesia, (3) the articles are written in English, (4) the articles indexed in SINTA, ranging

from SINTA 1 to SINTA 2, and (5) the articles published in the year of 2019-2021.

Lexical bundles in the corpus shown in Table 1 then were extracted using a corpus analysis toolkit namely Antconc 3.5.9 (Anthony, 2004). This software displays clusters of words based on the determined criteria and orders them alphabetically or even by frequency. For frequency analysis, this research focuses on 4-word bundles as an applicable criterion that is used to generate the most frequent lexical bundles that occur in both corpora. This is in line with Hyland (2008) who stated that 4-word lexical bundles are far more common than 5-word bundles and offer a clearer range of structures and functions than 3-word bundles. We also set a normalized frequency threshold with a minimum occurrence of 10 and a minimum distribution of 20 different texts. Lexical bundles with a high frequency are those that are used very often, while those with a low frequency are used less often.

Thenceforward, lexical bundles can be classified according to their structural patterns. Some common structural patterns include: noun-based bundles, prepositional-based bundles, and verb-based bundles (Biber et al.,1999) . Noun-based bundles consist of a noun and a modifier, such as "*the main point*" or "*a number of factors*". Prepositional-based bundles consist of a preposition and a noun, such as "*in the end*" or "*on the other hand*". While verb-based bundles consist of a verb and a complement, such as "*to make a point*" or "*to come up with a solution*." Besides structural classifications, lexical bundles also can be classified according to the functional classifications. The functional classifications of lexical bundles in this present research consist of research-oriented, text-oriented, and participant-oriented bundles (Hyland, 2008; Salazar, 2014). The subcategories of research-oriented bundles are: location, procedure bundles, quantification, description, and topic. This function help writers to

structure their activities and experiences of the real world. The second function is text-oriented bundles that concerned with the organization of the text and its meaning as a message or argument. The subcategories are: transition signals, resultative signals, and framing signals. Last function is participant-oriented bundles which subcategories are: stance features and engagement features. This function focused on the writer or reader of the text.

3. FINDINGS AND DISCUSSION

The frequency of a lexical bundle is the number of times it occurs in a corpus of text. In general, lexical bundles with a high frequency are those that are used very often in a particular language. The frequency of lexical bundles can also be affected by the length of the bundle. Shorter bundles are more frequent than longer bundles because shorter bundles are easier to remember and use, and they are less likely to be interrupted by other words or phrases. The tables below demonstrated the most frequent 4-word bundles in both natural and social science corpora.

Table 2. The most frequent lexical bundles in natural science research articles

No	Natural Science	Frequency
1	the results of the	195
2	in the form of	142
3	can be seen in	96
4	in the learning process	90
5	the results of this	76
6	results of this study	69
7	be seen in table	57
8	in this study were	55
9	is one of the	55
10	to be able to	53
11	based on the results	52
12	on the results of	52
13	in this study the	50
14	in this study was	49
15	can be used as	48
16	used in this study	48
17	in line with the	45

18	it can be concluded	44
19	that there is a	44
20	can be concluded that	43
21	of this study was	39
22	be concluded that the	36
23	be used as a	36
24	is in line with	36
25	the results showed that	35
26	this study aims to	34
27	in accordance with the	33
28	on the other hand	33
29	it is necessary to	32
30	can be used to	31
31	of this study is	26
32	the result of the	26
33	this study aimed to	25
34	results and discussion the	24
35	this study was to	22
36	the purpose of this	21

Table 3. The most frequent lexical bundles in social science research articles

No	Social Science	Frequency
1	in the form of	241
2	on the other hand	116
3	in the context of	85
4	is one of the	78
5	the results of the	73
6	at the same time	59
7	to be able to	56
8	is in line with	55
9	in line with the	54
10	as one of the	48
11	as well as the	47
12	can be seen from	40
13	the other hand the	40
14	it can be seen	39
15	the end of the	37
16	in this study the	35
17	the result of the	34
18	the use of the	34
19	at the end of	32
20	the context of the	32
21	can be seen in	28
22	the results of this	27
23	in addition to the	26
24	as a result of	25
25	one of the most	25

26	this study aims to	25

The most frequent lexical bundle in both natural science and social science is "*the results of the*". This is followed by "*in the form of*" in natural science and "*on the other hand*" in social science. These lexical bundles are often used to introduce the results of a study or to introduce a contrast between two ideas. Other common lexical bundles in natural science include "*can be seen in*", "*in the learning process*", and "*results of this study*". These lexical bundles are often used to describe the findings of a study or to discuss the implications of the findings. Common lexical bundles in social science include "*in the form of*, "*on the other hand*", and "*in the context of*". These lexical bundles are often used to introduce a concept or idea, to contrast two ideas, or to provide a context for the discussion.

The tables show that there are some lexical bundles that are more common in natural science than in social science, and vice versa. For example, the lexical bundle "*in the learning process*" is more common in natural science, while the lexical bundle "*on the other hand*" is more common in social science. This difference in the frequency of lexical bundles can be explained by the different research questions and methods that are used in natural science and social science RA. Natural science RA are often focused on the physical world, while social science research is often focused on human behavior. This difference in focus leads to different ways of thinking and writing about research, which is reflected in the different lexical bundles that are used.

Table 4: The structural forms of the lexical bundles in natural science research articles

Structural forms		Types		Lexical bundles
Noun-based	NP with of- phrase fragment	4	1	the results of the
			2	the results of this
			3	the result of the
			4	the purpose of this
	NP with other post-modifier fragments	3	5	the results showed that
			6	this study aims to
			7	this study aimed to
	Pronoun/NP + be + (…)	1	8	this study was to
	Total	8		
Prepositional-based	PP with embedded of-phrase fragment	5	9	in the form of
			10	to be able to
			11	on the results of
			12	of this study was
			13	of this study is
	Other PP (fragment)	7	14	in the learning process
			15	in this study were
			16	in this study the
			17	in this study was
			18	in accordance with the
			19	on the other hand
			20	in line with the
	Total	12		
Verb-based	Anticipatory it + VP/AP	2	21	it can be concluded
			22	it is necessary to
	Passive verb	8	23	can be seen in
			24	be seen in table
			25	based on the results
			26	can be used as
			27	used in this study
			28	can be concluded that
			29	be used as a
			30	can be used to
	Be + NP/AP	1	31	is in line with
	(VP +) that- clause fragment	3	32	that there is a
			33	is one of the
			34	be concluded that the
	Other expressions	2	35	results of this study
			36	results and discussion the
	Total	16		

The data exposed on the table 4 indicates that the mostly applied structural forms of the lexical bundles in natural science RA is verb-based with a total of 16 lexical bundles used in the corpus. The form of passive verb such as *can be seen in* is more frequently used compared to other forms with a total of 8 times. Biber et al (1999) describes that passive verb are useful in identifying tabular/graphic display of data. And identifying the basis of some of some finding or assertion. In other words, the passive words structural forms guide the readers to focus on the research rather than the writer's point of view. Furthermore, the other Verb-Based structural forms identified in the corpus are (VP +) that- clause fragment with 3 times used, Anticipatory it + VP/AP with 2 times used, other expressions with 2 times used be + NP/AP with 1 time used. The next lexical bundles with high frequent usage after the verb-based structural form is prepositional-based with a total of 12 lexical bundles used in the articles with the Other PP (fragment) type such as *in the learning process,* and *on the other hand* used 7 times and PP with embedded of-phrase fragment used 5 times. The least lexical bundle that is identified in the articles is Noun-based structural form. This lexical bundle is often used

for describing, identifying of place, size, and amount. There are 8 lexical bundles that follow noun-based structural form. In addition, the articles deliver 4 NP with of- phrase fragment, 3 NP with other post-modifier fragments, and 1 pronoun/NP + be + (...). It can be concluded that in natural science corpus, the verb-based is mostly utilized since in this type of corpus, tabulation and graphic are presented to show the data in the corpus. The verb-based lexical bundle is helpful to deliver the importance of data shown in tabulation and graphic. Moreover, the use of passive verb is mostly encouraged by the fact that the writing style of a research article mostly use a formal writing style where the object from a sentence is more highlighted and emphasized rather than using first person subject as the beginning of the sentence.

Table 5: The structural forms of the lexical bundles in social science research articles

Structural forms		Types		Lexical bundles
Noun-based	NP with of- phrase fragment	7	1	the results of the
			2	the end of the
			3	the result of the
			4	the use of the
			5	the context of the
			6	the results of this
			7	one of the most
	NP with other post-modifier fragments	2	8	the other hand the
			9	this study aims to
	Total	9		
Prepositional-based	PP with embedded of-phrase fragment	6	10	in the form of
			11	in the context of
			12	to be able to
			13	as one of the
			14	as a result of
			15	at the end of
	Other PP (fragment)	5	12	on the other hand
			13	at the same time
			14	in line with the
			15	in this study the
			17	in addition to the
	Total	11		
Verb-based	Anticipatory it + VP/AP	1	18	it can be seen
	Passive verb	2	19	can be seen from
			20	can be seen in
	Be + NP/AP	2	21	is one of the
			22	is in line with
	Adverbial clause	1	23	as well as the
	Total	6		

The data presented in table 5 indicates reveals that the mostly applied structural forms of the lexical bundles in social science RA is prepositional-based with 11 lexical bundles from a total of 23 lexical bundles identified in the articles. Biber et al. (1999) explains that prepositional-based frequently used to mark temporal relations, contradict and compare data. PP with embedded of- phrase fragment type such as *in the*

form of, and *in the context of* dominates the use of this type. Other type of this structural form which is other PP (fragment) such as *on the other hand* can also be found from the articles. The next frequently used lexical bundle in this type of corpus is Noun-based lexical bundle. The type of this bundle is presented 9 times. The first noun-based lexical bundle type is NP with of- phrase fragment such as *the results of the*. This type is presented 7 times. In addition, the other type of noun-Based lexical bundle is NP with other post-modifier fragments such as *this study aims to*. This type is appeared two times in the corpus. The least lexical bundle that is established in the articles is verb-based structural form. The corpus presents this bundle 6 times. In social science corpus, arguments, comparisons, and contrast ore often presented. Regarding with the fact that prepositional-based lexical bundle is highly used in this corpus, (Yuliawati et al., 2020) stated that prepositional-based bundle commonly relate to the text structure and its meaning, particularly to established arguments by describing limiting conditions.

Table 6: Functional classification of lexical bundles in natural and social science research articles

Function	Natural Science	Social Science
	Types	Types
Research-oriented bundles	15	12
Location	5	4
Procedure	1	1
Quantification	1	3
Description	6	4
Topic	2	-
Text-oriented bundles	12	10
Transition signals	4	6
Resultative signals	3	1
Structuring signals	-	-
Framing signals	5	3
Participant-oriented bundles	9	4
Stance features	1	1
Engagement features	8	3

As mentioned in the method section, the functional classifications of lexical bundles in this recent study refers to the theory proposed by Hyland (2008 & 2012). The results expose that between Natural science

RA and Social Science RA have similarities and differences. The first similarities from both research is that both of them has the same order in terms of the frequency of the functional use. What is meant by this is that both RA mostly apply research-oriented bundles and then the text-oriented bundles, and the last is the participant-oriented bundles.

Regarding with the differences, it can be shown from the table that the natural science RA has more frequency of the function with the total of 36 function while the social science use 26 function. Furthermore, by looking this data the natural science RA has more data to be displayed and explained as it consists of more lexical bundles.

In regards with the first functional classification which both of the RA frequently used, In the research-oriented bundles classification, the natural science RA applies 15 times, on the other hand, the social science applies the function 12 times. In natural science RA, the description type is displayed 6 times, location type is utilized 5 times, the topic is taken advantage 2 time, and both the procedure and quantification type are used 1 time. Furthermore, in social science RA, both location and description type are applied 4 times, the quantification type is used 3 times, the procedure is applied one time. However, in social science RA, it can be assured that this study doesn't identify topic type usage.

In addition with the second functional classification, the similarity between both RA can be seen from the table that both of them do not use the type of structuring signals. Moreover, the most frequent used type of the second functional classification in the natural science RA is the framing signals which is used 5 times. In contrast, the social science use the transition signals as the most frequent type. In total, the natural science RA applies text-oriented bundles 12 times with the rest of the types which are 3 the transition signals 4 resultative signals.

The last functional classification of lexical bundles is the

participant-oriented bundles. This is the least classification which is used by both of the RA. Both of the RA similarly use the engagement features in a higher number compare to the stance features. The Natural science RA use 8 engagement feature while on the other side, the social science RA use 3 engagement feature. Interestingly, both of the RA only use the stance features one time. It can be assumed that both of the RA avoid writer's attitude and evaluations.

4. Conclusions

The better your paper looks, the better the Journal looks. Thanks for your cooperation and contribution.

The present research explores the patterns of lexical bundles in the corpora of natural and social sciences. The corpus consists of 200 research articles, compiling almost 1.000 words which is a large collection of text. Hence, a corpus tool namely Antconc is used to analyze the frequency, structure, and function of lexical bundles in different disciplinary areas.

The most frequent lexical bundle in both natural science and social science is "*the results of the*." This is followed by "*in the form of*" in natural science and "*on the other hand*" in social science. These lexical bundles are often used to introduce the results of a study or to introduce a contrast between two ideas. The frequency of lexical bundles can vary between natural science and social science RA. For example, the lexical bundle "*in the learning process*" is more common in natural science, while the lexical bundle "*on the other hand*" is more common in social science. This difference in the frequency of lexical bundles can be explained by the different research questions and methods that are used in natural science and social science RA. The different form of lexical bundles used in natural science and social science RA reflect the different ways of thinking and writing about research in these two disciplines.

In terms of structural forms, the most common structural form of lexical bundles in natural science RA is

verb-based. While the most common structural form of lexical bundles in social science RA is prepositional-based. The contradiction suggests that in terms of structure, there are differences between natural and social science RA writing styles, how the authors establish arguments, and also present results.

Meanwhile from the functional classifications, we can conclude that both types of RA have the same order in terms of the frequency of the functional use, with research-oriented bundles being the most frequent, followed by text-oriented bundles, and then participant-oriented bundles. However, there are some differences between the two types of RA, with natural science RA having a higher frequency of function use than social science RA. The differences in functional use between natural science and social science RA can be explained by the different research questions and methods that are used in these two disciplines. Natural science RA are often focused on the physical world, while social science research is often focused on human behavior. This difference in focus leads to different ways of thinking and writing about research, which is reflected in the different functional uses of lexical bundles.

Finally, analyzing lexical bundles in different disciplinary areas can be helpful for a number of reasons. First, it can help to identify the most common ways of communicating in a particular discipline. This information can be used to improve the readability of research articles in one discipline, as well as to help researchers learn the conventions of writing in one discipline. Second, analyzing lexical bundles can help to identify the differences in how communication is structured in different disciplines. This information can be used to improve the cross-disciplinary communication of research findings. Third, analyzing lexical bundles can help to identify the evolution of language use in a particular discipline. This information can be used to track the development of new research methods and theories, as well as to

identify the impact of changes in the social and political landscape on the way that research is communicated. Overall, analyzing lexical bundles in different disciplinary areas can be a valuable tool for improving the communication of research findings, as well as for understanding the evolution of language use in different disciplines.

References

Anthony, L. (2004). AntConc: A Learner and Classroom Friendly, Multi-Platform Corpus Analysis Toolkit. *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, 7–13.

Baker, P., & Chen, Y.-H. (2010). Lexical Bundles in L1 and L2 Academic Writing. *Language Learning and Technology*, *14*(2), 30–49. http://llt.msu.edu/vol14num2/chenbaker.pdf

Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan; Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman. https://doi.org/10.1177/0075424202250290

Budiwiyanto, A., & Suhardijanto, T. (2020). Indonesian lexical bundles in research articles: Frequency, structure, and function. *Indonesian Journal of Applied Linguistics*, *10*(2), 292–303. https://doi.org/10.17509/ijal.v10i2.28592

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, *27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics*, *32*, 150–169. https://doi.org/10.1017/S0267190512000037

Kashiha, H., & Heng, C. S. (2013). An exploration of lexical bundles in academic lectures: Examples from hard and soft sciences. *Journal of Asia TEFL*, *10*(4), 133–161.

Kwary, D. A., Ratri, D., & Artha, A. F. (2017). Lexical bundles in journal articles across academic disciplines. *Indonesian Journal of Applied Linguistics*, *7*(1), 132–140. https://doi.org/10.17509/ijal.v7i1.6866

Salazar, D. (2014). *Lexical Bundles in Native and Non-native Scientific Writing* (Studies in). John Benjamins Publishing Company.

Thanh Tuyen, K., Osman, S. Bin, Cong Dan, T., & Shafrin Binti Ahmad, N. (2016). Developing Research Paper Writing Programs for EFL/ESL Undergraduate Students Using Process Genre Approach.

Higher Education Studies, *6*(2), 19. https://doi.org/10.5539/hes.v6n2p19

Yuliawati, S., Ekawati, D., & Mawarrani, R. E. (2020). *Penulisan Akademik: Perspektif Linguistik Korpus dan Analisis Wacana* (N. Darmayanti (ed.)). Unpad Press.

Author **Annur Karima Zulyanputri,** Student at the Linguistics Study Program, Padjajaran Universtity.