https://doi.org/10.5392/IJoC.2018.14.4.070

Lexical Bundles in Computer Science Research Articles: A Corpus-Based Study

Je-Young Lee
Department of English Education
Jeonju University, Jeonju-si, Jeollabuk-do, 55069, South Korea

Hye Jin Lee
College of Liberal Arts
Wonkwang University, Iksan-si, Jeollabuk-do, 54538, Korea

ABSTRACT

The purpose of this corpus-based study was to find 4-word lexical bundles in computer science research articles. As the demand for research articles (RAs) for international publication increases, the need for acquiring field-specific writing conventions for this academic genre has become a burning issue. Particularly, one area of burgeoning interest in the examination of rhetorical structures and linguistic features of RAs is the use of lexical bundles, the indispensable building blocks that make up an academic discourse. To illustrate, different academic discourses rely on distinctive repertoires of lexical bundles. Because lexical bundles are often acquired as a whole, the recurring multi-word sequences can be retrieved automatically to make written discourse more fluent and natural. Therefore, the proper use of rhetorical devices specific to a particular discipline can be a vital indicator of success within the discourse communities. Hence, to identify linguistic features that make up specific registers, this corpus-based study examines the types and usage frequency of lexical bundles in the discipline of CS, one of the most in-demand fields world over. Given that lexical bundles are empirically-derived formulaic multi-word units, identifying core lexical bundles used in RAs, they may provide insights into the specificity of particular CS text types. This will in turn provide empirical evidence of register specificity and technicality within the academic discourse of computer science. As in the results, pedagogical implications and suggestions for future research are discussed.

***Key words**: Lexical Bundles, Computer Science, Research Articles, Corpus, English for Academic Purposes.*

1. INTRODUCTION

Research articles (RAs) published in prestigious international journals are an indication of scholarly productivity and achievement. In this context, there is a worldwide demand for high-quality RAs and it led to a growing interest in academic writing among researchers. Accordingly, research on RAs has gained momentum with a particular attention being paid to the illumination of RA linguistic features and rhetorical conventions across different disciplines. Along this line, a growing number of studies have found that there are formulaic multi-word combinations, or lexical bundles, which play a fundamental role in framing academic discourses in RAs. Lexical bundles are frequently occurring word sequences in a given register and they typically behave as single language units. Given that humans can hold and recall approximately four to seven chunks of information, lexical bundles can alleviate the cognitive loads of language processing while allowing writers to produce more credible academic voices. That is, awareness of lexical bundles, which endow interpretive frames for developing discourse in fulfilling genre expectations, can empower novice or non-native writers to surmount the challenge of finding the right words to express their expertise, and can thus help them to display greater processing efficiency.

Disciplines are distinguished by the shared features of their associated domains of inquiry and, in many instances, such variations can affect the ways in which knowledge is communicated. That is, different disciplines have idiosyncratic discourse preferences to fulfill communicative functions. Certain disciplines display marked choices in the use of lexical bundles that are distinguishable from one another. In this regard, lexical bundles can also bring to light whether different disciplines use distinctive sets of lexical bundles, and as such they have garnered considerable research attention to date. In particular, previous studies that examined the overlaps and

This is an excellent paper selected from the papers presented at 2018 Korea Contents Association Spring Academic Conference.

** Corresponding author, Email: lee.hyejin0419@gmail.com*

Manuscript received Nov. 26, 2018; revised Dec. 21, 2018; accepted Dec. 21, 2018

divergences of lexical bundles across a range of fields of study have identified disciplinary variations in the structural and functional application of lexical bundles. For example, some studies conducted a cross-disciplinary investigation to identify similarities and differences of lexical bundles across disciplines [1]-[3] while other studies identified structural and functional features of lexical bundles within one discipline such as applied linguistics [4], medical [5], chemistry [6], education [7], and telecommunications [8]. The shared conclusion drawn from the previous studies attest to the peculiarity and heterogeneity of bundles in professional academic discourse. They also accentuated the fact that the proper use of formulaic word combinations that are semantically and syntactically compositional can fulfill the rhetorical functions instrumental in RAs.

Although the scholarly endeavors have successfully advanced knowledge in this domain, surprisingly, little research has concentrated on the field of CS, one of the most promising and popular fields in the world. The growing global recognition and importance of the diverse fields of computer science has brought with it a corresponding and ever-increasing demand for publishable research articles to further scholarly exchange and international communication in this area. However, the outcomes of previous studies have yet to be validated and, thus, there are remaining questions as to the specificity of lexical bundles in the field of CS. In this regard, there is a need to develop a list of pedagogically valuable CS lexical bundles, which can lend insights into the phraseological features of specific language use in CS writing. To undertake a comprehensive analysis of lexical bundles embedded in computer science research articles (CSRAs), this exploratory corpus-driven study aims to identify the most recurrent multi-word sequences and to reveal the extent to which the lexical bundles achieve discourse functions within written repertoires of CS. Although lexical bundles (or n-grams) are of varying lengths, four-word lexical bundles are the most researched bundles in contemporary literature on the basis that they can be the most revealing in terms of text patterning. As such, to better contextualize the scope of the research, this paper examines the forms, structures, and functions of four-word bundles in a corpus of RAs in the discipline of CS. The two research questions that underpin this study are as follows:

1. What are the most common 4-word lexical bundles found in CSRAs?
2. What are the structural features of the 4-word lexical bundles in CSRAs?

2. MATERIAL AND METHOD

2.1 Corpus

In order to analyze lexical bundles in research articles in the field of computer science, the Computer Science Corpus (CSC) was compiled by the authors. It consisted of research articles written in English language from 27 SCIE-indexed computer science journals (listed in Table 1) published by Association for Computing Machinery (ACM). The ACM is one of the world-famous academic organizations in various computer science field, such as computer education, computer systems, hardware, information systems, mathematical computing, networks, and security/privacy. To maintain representativeness, five articles in each of 27 journals published between 2016 and 2018 were randomly selected and incorporated into the CSC.

Table 1. List of SCIE-indexed ACM journals included in the CSC

1. ACM Transactions on Algorithms (TALG)
2. ACM Transactions on Applied Perception (TAP)
3. ACM Transactions on Architecture and Code Optimization (TACO)
4. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)
5. ACM Transactions on Autonomous and Adaptive Systems (TAAS)
6. ACM Transactions on Computer Logic (TOCL)
7. ACM Transactions on Computer Systems (TOCS)
8. ACM Transactions on Computer-Human Interaction (TOCHI)
9. ACM Transactions on Computing Education (TOCE)
10. ACM Transactions on Database Systems (TODS)
11. ACM Transactions on Design Automation of Electronic Systems (TODAES)
12. ACM Transactions on Embedded Computing Systems (TECS)
13. ACM Transactions on Graphics (TOG)
14. ACM Transactions on Information Systems (TOIS)
15. ACM Transactions on Intelligent Systems and Technology (TIST)
16. ACM Transactions on Internet Technology (TOIT)
17. ACM Transactions on Knowledge Discovery from Data (TKDD)
18. ACM Transactions on Mathematical Software (TOMS)
19. ACM Transactions on Modeling and Computer Simulation (TOMACS)
20. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)
21. ACM Transactions on Privacy and Security (TOPS)
22. ACM Transactions on Programming Languages and Systems (TOPLAS)
23. ACM Transactions on Reconfigurable Technology and Systems (TRETS)
24. ACM Transactions on Sensor Networks (TOSN)
25. ACM Transactions on Software Engineering and Methodology (TOSEM)
26. ACM Transactions on Storage (TOS)
27. ACM Transactions on the Web (TWEB)

The size of the CSC was roughly 1.3 million words. Its size seemed appropriate when comparing other similar corpus. For example, Chen and Ge complied 1.9 million-word Whole Paper Corpus [9], Martínez, Beck, and Panza constructed 0.8 million-word AgroCorpus [10], and Jalali and Moini built 0.4 million word CIMRA (Corpus of Introduction section of Medical Research Articles) [5].

2.2 Software and Data Analysis

The computer software used in this study was AntConc 3.5.7 [11], a multi-platform freeware for corpus analysis. It provides various functions like showing concordance lines, analyzing collocation, creating word lists, and, above all, analyzing lexical bundles (N-grams). It was used for analyzing and counting 4-word lexical bundles in the CSC.

In this study, 4-word bundles that occurred more than 26 times (20 times per one million word) [5], [12] were regarded as the important ones in computer science field. In terms of range, lexical bundles occurred at least in 9 journals were selected. 4-word bundles are much more common than 5-word bundles and offer a clear range of structures and functions than 3-word bundles [2]. In addition, many 4-word bundles contain the structure of 3-word bundles [1].

3. RESULTS

3.1 Frequent Lexical Bundles in the CSC

As previously stated, the cut-off frequency of 26 was applied as criterion in the selection of 4-word lexical bundles. It yielded 137 different lexical bundles in the CSC. This number is higher than 62 in the fiction corpus of Allan [12] and slightly lower than 161 in medical research article corpus of Jalali and Moini [5]. You can see the 10 most frequent 4-word lexical bundles in the following Table 2, and the full list in Appendix 1.

Table 2. The 10 most frequent 4-word lexical bundles in the CSC

No.	Lexical Bundles	Frequency	Range
1	on the other hand	235	26
2	as shown in figure	234	23
3	in this article we	226	26
4	in this section we	188	27
5	in the case of	145	25
6	the size of the	131	23
7	the total number of	121	22
8	is the number of	117	25
9	can be used to	115	25
10	in the context of	109	24

3.2 Structure of Lexical Bundles in the CSC

In order to find out the structural features of the lexical bundles, the 60 most frequent 4-word lexical bundles were classified according to their structural characteristics, verb phrases (VPs), noun phrases (NPs), and prepositional phrases (PPs). As shown in Figure 1, the distribution of VP was 28%. It is slightly lower proportion than those of medical RAs and Fictions [12].

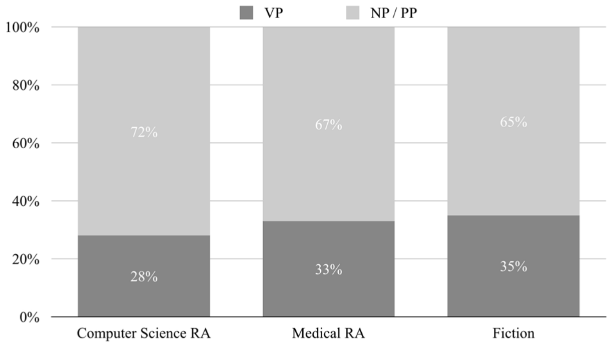

Fig. 1. Distribution of the 60 most frequent 4-Word Lexical Bundles in the CSC

4. CONCLUSIONS

Research articles, which serve as a core repository of field-specific knowledge, contain a set of disciplinary rhetorical conventions to facilitate knowledge exchange and to appeal to the interests of research communities. Thus, identifying the compositional features of lexical bundles in research articles based upon their discourse functions, structures, and frequency can be of value. To empirically examine lexical bundles, this study adopted a corpus-driven approach. The pedagogical implication of this approach for identifying lexical bundles is based on the assumption that most recurrent formulaic multi-word units are of the utmost currency and practicality and thus deserving of much pedagogical attention.

The analysis of this study suggested that the lexical bundles used in the written discourse of CS tended to follow pre-fabricated structures and employed numeric representations, which could be attributable to the abstract and mathematical nature of computer science rhetoric. To illustrate, with regard to the structural types of 4-word lexical bundles, NPs/PPs were more evident than VPs. The analysis also revealed that most pronounced functional aspects of lexical bundles are discourse organizers that incorporated noun and prepositional phrase fragments. For example, most lexical bundles served topic introduction functions (i.e., "in this article we," "in this section we") to offer overt signals that either a new topic or subtopic is being introduced as well as topic elaboration functions (i.e., "on the other hand") to supply more information to the topic. Furthermore, the findings also indicated that the CSC are dominated by referential lexical bundles that refer to size, amount, number, or quantity as in "the size of the," the total number of," "is the number of." This may be a consequence of the nature of computer science, which is based in mathematics, coding and algorithms [13].

Although this research paper has contributed to the current literature of lexical bundles, the findings put forward some possibilities for future research directions. First, the present study has investigated lexical bundles in the corpus of CSRAs as a whole. Future studies can build on this present research by identifying the use of lexical bundles in different sections of CSRAs (i.e., abstract, introduction, methods, results, discussion and conclusion) to see different repertoires of lexical bundles in each section. Furthermore, there is as yet no empirical evidence that indicates genre variations of lexical bundles in the discipline of CS. The corpus of the present study exclusively included CSRAs, one single genre so as to secure our empirical claims. However, further research on a larger scale by collecting data from other sub-disciplines of computer science may provide a more representative picture of lexical bundles used in this field. For example, future research can expand this study by comparing the use of lexical bundles in different genres of CS such as textbooks, grant proposals, textbooks, conference papers, Ph. D and master's theses, book chapters, etc. Future research that goes beyond a single-genre approach may offer instructional guidance on which lexical bundles language learners need to learners need to target in the milieu of academic learning and professional career development.

Lastly, future research endeavors can be directed towards comparing the use of lexical bundles produced by first/native language (L1)-English and second language (L2)-English writers with varying degrees of expertise. Humans have a limited cognitive capacity for processing information, which indicates that storing formulaic multi-word chunks or lexical bundles can allow more prompt language comprehension and production. Regardless of their language backgrounds, all novice L1 and L2 writers are expected to be conversant in the use of discipline-specific discourse conventions, especially lexical bundles. While the task can be challenging for both L1 and L2 writers alike, previous studies have revealed that L1 writers get a head start due to their familiarity with conventional lexical bundles. Native speakers' extensive exposure to their mother tongue over a number of years results in the intuitive use of common word combinations and their mental inventory of these bundles enable them to bypass the processing route by which they are retrieved. In other words, L1 writers, indeed, access processing benefits through their mastery of formulaic language sequences or lexical bundles. Thus, gaining a deeper insight into how these L1 and L2 writers use bundles differently from structural and functional aspects can provide useful pedagogical implications for vocabulary instructions.

The existing body of literature in general highlighted the significance of possessing a solid command of formulaic language sequences to promote language proficiency and professional collegiality within target discourse communities. In this regard, the list of corpus-informed lexical bundles with the structural and functional classifications identified in this study can be of pedagogical value for vocabulary instruction, especially in the field of computer science.

REFERENCES

[1] V. Cortes, "Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology," English for Specific Purposes, vol 23, no. 4, 2004, pp. 397-423.

[2] K. Hyland, "As Can Be Seen: Lexical Bundles and Disciplinary Variation," English for Specific Purposes, vol. 27, no. 4, 2008, pp. 4-21.

[3] T. Omidian, H. Shahriari, and A. Siyanova-Chanturia, "A Cross-disciplinary Investigation of Multi-word Expressions in the Moves of Research Article Abstracts," Journal of English for Academic Purposes, vol. 36, 2018, pp. 1-14.

[4] M. Kazemi, M. Kohandani, and N. Farzaneh, "The Impact of Lexical Bundles on How Applied Linguistics Articles are Evaluated," Procedia: Social and Behavioral Sciences, vol. 98, 2014, pp. 870-875.

[5] Z. S. Jalali and M. R. Moini, "Structure of Lexical Bundles in Introduction Section of Medical Research Articles," Procedia-Social and Behavioral Sciences, vol. 98, 2014, pp. 719-726.

[6] L. Valipoor, *A Corpus-based Study of Words and Bundles in Chemistry Research Articles,* MA thesis, University of Kashan, 1998.

[7] N. Parvizi, *Identification of Discipline-specific Lexical Bundles in Education,* MA thesis, University of Kashan, 2011.

[8] F. Pan, R. Reppen, and D. Biber, "Comparing Patterns of L1 versus L2 English Academic Professionals: Lexical Bundles in Telecommunications Research Journals," Journal of English for Academic Purposes, vol. 21, 2016, pp. 60-71.

[9] Q. Chen and G. Ge, "A Corpus-based Lexical Study on Frequency and Distribution of Coxhead's AWL Word Families in Medical Research Articles (RAs)," English for Specific Purpose, vol. 26, no. 4, 2007, pp. 502-514.

[10] I. A. Martínez, S. C. Beck, and C. B. Panza, "Academic Vocabulary in Agriculture Research Articles: A Corpus-based Study," English for Specific Purposes, vol. 28, no. 3, 2009, pp. 183-198.

[11] L. Anthony, Antconc 3.5.7: A Free Text Analysis Software. Available on line at http://www.antlab.sci.waseda.ac.jp, 2018.

[12] R. Allan, "Lexical Bundles in Graded Readers: To What Extent Does Language Restriction Affect Lexical Patterning?," System, vol. 59, 2016, pp. 61-72.

[13] V. C. Galpin and I. D. Sanders, "Perceptions of Computer Science at a South African University," Computers & Education, vol. 49, no. 4, 2007, pp. 1330-1356.

Je-Young Lee

He received the B.A., M.A., and Ed.D. degree in English Education from Korea National University of Education in 2000, 2004, 2013 respectively. He was the assistant professor in Sehan University from 2014 to 2017. In 2017, he started working in Jeonju University as the assistant professor of Dept. of English Education. His main research interests are corpus linguistics, technology-enhanced language learning, and research synthesis.

Hye Jin Lee
She obtained her B.A. in English Education from Wonkwang University, Korea in 2011 and M.A. in TESOL from Oklahoma City University, the United States in 2013. She received her Ph.D. in Foreign and Second Language Education from the State University of New York at Buffalo in 2016. Since 2017, she has been a professor of the College of Liberal Arts at Wonkwang University. Her main research interests include corpus linguistics, lexical bundles, statistical linguistics, TESOL, English for specific purposes and disciplinary writing.

APPENDIX

Full list of 4-word lexical bundles in the CSC

No.	Lexical Bundles	Frequency	Range
1	on the other hand	235	26
2	as shown in figure	234	23
3	in this article we	226	26
4	in this section we	188	27
5	in the case of	145	25
6	the size of the	131	23
7	the total number of	121	22
8	is the number of	117	25
9	can be used to	115	25
10	in the context of	109	24
11	with respect to the	106	24
12	the performance of the	99	22
13	if and only if	94	14
14	as well as the	88	23
15	as the number of	88	20
16	a large number of	86	22
17	to the number of	86	20
18	of the number of	85	21
19	is shown in figure	82	18
20	at the same time	81	23
21	and the number of	78	22
22	at the end of	73	17
23	the rest of the	64	23
24	on the number of	64	21
25	with the number of	64	16
26	it is possible to	61	20
27	shown in figure the	60	22
28	the results of the	60	18
29	the end of the	60	17
30	in the number of	60	16
31	in this case the	59	20
32	the quality of the	58	22
33	in terms of the	57	24
34	is based on the	56	22
35	it is important to	56	20
36	in this work we	56	19
37	can be found in	55	19
38	in the form of	53	15
39	a small number of	52	19
40	to the fact that	52	21
41	in the next section	51	24
42	as discussed in section	51	15
43	is the set of	51	15
44	the best of our	49	23
45	the other hand the	49	17
46	is due to the	48	22
47	in addition to the	48	20
48	are shown in figure	48	18
49	when the number of	48	16
50	the number of iterations	48	11
51	best of our knowledge	47	23
52	to the best of	47	23
53	as a function of	47	17
54	we can see that	47	16
55	is organized as follows	45	20
56	article is organized as	44	20
57	for each of the	44	18
58	we assume that the	44	17
59	a wide range of	43	19
60	the maximum number of	43	16
61	that there is a	43	15
62	the length of the	42	15
63	as described in section	41	21
64	the set of all	41	13
65	the average number of	41	12
66	of a set of	40	16
67	as a result of	40	15
68	the fact that the	39	19
69	due to the fact	38	19
70	is similar to the	38	19
71	that the number of	38	19
72	as shown in table	38	16
73	in other words the	38	16
74	as a result the	38	14
75	in the presence of	38	13
76	on the basis of	38	11
77	is defined as the	37	17
78	reduce the number of	37	15
79	by the number of	36	18
80	that can be used	36	17
81	the state of the	36	15
82	the reason is that	36	14
83	is defined as follows	36	13
84	this is due to	35	19
85	in the following we	35	12
86	is one of the	34	18
87	the complexity of the	34	18
88	the same number of	34	15
89	the accuracy of the	33	15
90	evaluate the performance of	33	12
91	we focus on the	32	17
92	it should be noted	32	14
93	in the worst case	32	12
94	the number of times	32	12

95	is a set of	32	11
96	a subset of the	31	18
97	the effectiveness of the	31	16
98	the impact of the	31	16
99	as illustrated in figure	31	15
100	we only need to	31	14
101	as can be seen	31	13
102	the values of the	31	13
103	are shown in table	30	16
104	than or equal to	30	14
105	with the help of	30	13
106	can be seen in	30	11
107	one of the most	29	18
108	we are interested in	29	17
109	the sum of the	29	15
110	can be applied to	29	14
111	in the same way	29	13
112	to this end we	29	13
113	the article is organized	28	18
114	as mentioned in section	28	15
115	for example in the	28	15
116	is the same as	28	15
117	rest of the article	28	15
118	the goal is to	28	14
119	to be able to	28	14
120	in the sense that	28	12
121	should be noted that	28	12
122	is said to be	28	9
123	is based on a	27	16
124	in the rest of	27	15
125	the structure of the	27	12
126	be a set of	27	9
127	at the beginning of	26	15
128	of this article is	26	15
129	it is easy to	26	14
130	the difference between the	26	14
131	the case of a	26	13
132	if there is a	26	12
133	our goal is to	26	12
134	is worth noting that	26	11
135	it is worth noting	26	11
136	the ratio of the	26	11
137	the behavior of the	26	10