Chabahar Maritime University
Iranian Journal of English for Academic Purposes ISSN: 2476-3187
IJEAP, 2021, 10(2), 90-105 (Previously Published under the Title: Maritime English Journal)

Lexical Bundles in Research Articles in Chemistry: A Structural Analysis

[1]Javad Zare*
[2]Leila Valipouri

Research Paper IJEAP- 2103-1697 DOR: 20.1001.1.24763187.2021.10.2.7.3
Received: 2021-04-16 Accepted: 2021-07-02 Published: 2021-07-04

Abstract

This paper reports the results of a corpus-driven study that investigated the structural frequency and diversity of lexical bundles in chemistry research articles. The investigation was based on a corpus of 1,185 chemistry research articles, totaling four million words. WordSmith was used to generate four-word lexical bundles and their concordance lines. Altogether, 223 lexical bundles were found. More than 55% of these bundles were phrasal; 35% of them were clausal. 'Prepositional phrases + of' bundles were the most frequently used bundles. Bundles with a passive verb, followed by a prepositional phrase fragment were the second most frequent bundles in the entire corpus. Bundles with the structure 'noun phrase + of' were the third and second most frequent bundles in the corpus and among phrasal bundles, respectively. In terms of diversity, bundles with 'noun phrase + of' and 'passive + prepositional phrase fragment' were the most varied and bundles with 'pronoun/noun phrase + be' structure were the least varied bundles in the entire corpus. Prepositional phrase bundles outnumbered noun phrase bundles in terms of frequency, whereas noun phrase bundles outnumbered prepositional phrase bundles in diversity. Altogether, the results show that frequency and diversity correlate with the type of phrasal lexical bundle. Moreover, the study suggests that different discourses are associated with different sets of lexical bundles with different frequency and diversity, due to the different communicative functions they follow. The paper ends with implications for future EAP research, materials development, and pedagogy.

Keywords: Lexical bundles; Research articles; Structural analysis; Chemistry research articles

1. Introduction

Research shows that language is formulaic in nature (e.g., Wray, 2002). Besides, it is claimed that awareness of formulas, i.e. recurrent multi-word combinations, facilitates language learning and leads to successful language production (Conklin & Schmitt, 2012). Research also shows that phrases are learned as unanalyzed wholes or chunks rather than individual words and that learning relies heavily on these expressions in the early stages of language acquisition (e.g., Ellis, 2002; Wray, 2002; Staples, Egbert, Biber & McClair, 2013). Some even equate frequent use of appropriate formulaic sequences or lexical bundles with language development and their absence with "lack of mastery of a novice writer in a specific disciplinary community" (Li & Schmitt, 2009, p. 86) (e.g., Ellis, 1996; Ellis & Simpson-Vlach, 2008). As "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber, Johansson, Leech, Conrad & Finegan, 1999, p. 990) that "show a statistical tendency to co-occur" (Biber & Conrad, 1999, p. 183), lexical bundles have received extensive attention in research studies (e.g., Biber et al., 1999; Hyland, 2008a, b; Zare & Naseri, 2020). Research shows that "a distinct set of lexical bundles, associated with [its] typical communicative purposes" is peculiar to a certain genre or register (e.g., Biber & Barbieri, 2007, p. 265). More importantly, bundles are reported to be associated with disciplines (e.g., Cortes, 2004; Hyland, 2008a). Hence, discipline-specific studies of lexical bundles are required. Although the literature abounds with studies of the use of lexical bundles in the discourse of different disciplines (e.g., Cortes, 2013; Coxhead & Byrd, 2010; Hyland, 2008a, b), to the best of our knowledge, no prior study has investigated the structural features of lexical bundles in chemistry research articles. To this end, the present study aimed to

[1] Assistant Professor of Applied Linguistics (Corresponding Author), javadzare@gmail.com; Kosar University of Bojnord, Bojnord, Iran.
[2] M.A. in TEFL, leila_valipoor_g@yahoo.com; University of Kashan, Kashan, Iran.

investigate the frequency and diversity of the structural patterns of lexical bundles in chemistry research articles (RAs).

2. Review of the Related Literature

Lexical bundles are defined as “words which follow each other more frequently than expected by chance, helping to shape text meanings and contributing to our sense of distinctiveness in a register” (Hyland, 2008a, p. 5). According to Biber et al. (1999), lexical bundles are sequences of three or more words that frequently recur in a genre or register and are usually identified with the use of a computer program. Typically, bundles are not idiomatic in meaning (Biber, Conrad & Cortes, 2003), and do not lend themselves to complete syntactic units (Biber & Conrad, 1999; Biber et al., 1999; Stubbs & Barth, 2003), although they correlate with certain structural categories that are used for classifying them.

Generally, there are three parameters that are taken into consideration when identifying lexical bundles: length, frequency, and dispersion. Length is the number of words each lexical bundle consists of. Three to six-word lexical bundles are usually investigated in studies. Four-word bundles are the most common types of bundles because they are more frequent and varied than other bundles (Hyland, 2012). Frequency is the number of times a sequence needs to occur to be considered as a lexical bundle. Different research studies set different frequencies from 10 up to 40 times per million words in a corpus. Dispersion is the number of times a bundle needs to occur in texts. Different studies set different dispersions, ranging from three to five. Dispersion is computed to make sure that the bundle is typical of the entire corpus (Hyland, 2012).

Studies of lexical bundles have typically explored functional and structural properties of bundles. In terms of function, Hyland (2008a, b) classified lexical bundles into research-oriented, text-oriented, and participant-oriented. Research-oriented bundles are lexical bundles that deal with real-world activities, including location, procedure, quantification, description, and topic bundles. Text-oriented bundles are used to express the organization of the text, including transition, resultative, structuring, and framing signal bundles. Participant-oriented bundles are lexical bundles that turn to the reader or writer, including stance and engagement bundles. Several studies have investigated the functional features of lexical bundles which, due to limit in word count, fall out of the scope of this paper (e.g., Esfandiari & Moein, 2016; Hyland, 2008a, b; Jalali & Moini, 2018; Zare & Naseri, 2020).

In terms of structure, investigating the formal features of lexical bundles is a frequent theme in prior research (e.g., Allan, 2016; Biber et al., 1999; Biber, Conrad & Cortes, 2004; Cortes, 2013; Hong, 2019; Hyland, 2008a, b; Jablonkai, 2010; Jalali & Moini, 2018; Jalali, Moini & Alaee Arani, 2015; Pan et al., 2016; Pérez-Llantada, 2014; Qin, 2014; Rahimi Azad & Modarres Khiabani, 2018; Ruan, 2016; Zare & Naseri, 2020). Biber et al. (1999) made a distinction between phrasal (e.g., in the case of) and clausal bundles (e.g., I don't know what). Phrasal bundles may incorporate noun phrases (e.g., the nature of the) and prepositional phrases (e.g., in the context of). Clausal bundles, on the other hand, may incorporate a simple verb phrase (e.g., have a look at) or a main clause (e.g., I don't know how). Cortes (2013) divided four-word and longer lexical bundles in research article introductions from different disciplines into four main groups: (1) lexical bundles that consist of noun phrase or prepositional phrase fragments (e.g., in the present study); (2) lexical bundles that consist of verb phrase fragments (e.g., little is known about); (3) lexical bundles that include dependent clause fragments (e.g., little is known about); and (4) lexical bundles that incorporate both noun and verb phrases (e.g., the objective of this study was to evaluate). In another study, Biber et al. (2004) compared lexical bundles in university classroom teaching and text books with bundles in conversation and academic prose and divides them into Type 1, Type 2, and Type 3. Type 1 lexical bundles include verb phrase fragments (e.g., it’s going to be); Type 2 bundles incorporate dependent clause fragments besides verb phrase fragments (e.g., I want you to); and Type 3 bundles consist of noun phrases (e.g., the end of the), and prepositional phrases (e.g., of the things that). In general, Type 1 and Type 2 bundles are clausal, whereas Type 3 bundles consist of phrasal components. While lexical bundles in conversations incorporate verb phrases and clause fragments, bundles in academic prose incorporate mostly noun phrase and prepositional phrase fragments. The lexical bundles identified in textbooks consist mostly of noun phrase and prepositional phrases. Bundles in university classroom teaching, however, reflect the characteristics of both written and spoken modes of language use, as they consist mostly of both clausal and phrasal fragments. In another study, Hyland (2008b) examined variation of lexical bundles in form, function, and structure in a corpus of 3.5 million words comprising the three genres of doctoral dissertations, master's theses, and research articles in four disciplines, i.e. electrical engineering, microbiology, business, and applied linguistics. Hyland found most of the bundles in the corpus to be parts of noun phrases or prepositional phrases and to end with prepositions, articles, and complimentizers. Structural analysis of the bundles showed that "several of these structures reflect the cautious limitations of academic discourse, typically through post-nominal modification, agent-evacuated passives and anticipatory-it patterns" (2008b, p. 48). In another study, Pan et al. (2016) investigated the use of lexical bundles by L1-English versus L2-English academic professionals in Telecommunications research papers and found that the lexical bundles that L2 writers use are mostly bundles with verbs and clause fragments (especially passive verb structures). On the other hand, the lexical bundles that L1 writers use mostly consist of noun phrases and prepositional phrases.

Generally, research on lexical bundles can be categorized into three groups: lexical bundles across proficiency levels, lexical bundles across disciplines, and lexical bundles across genres. Regarding lexical bundles across proficiency level, research shows that the use of lexical bundles varies across different proficiency levels. For example, Staples et al. (2013) investigated the use of lexical bundles in the written responses of learners with different proficiency levels and found that lower-level English learners use more lexical bundles. In other words, learners tend to use fewer bundles, as they gain proficiency in English. In another study, Ädel and Erman (2012) compared the use of lexical bundles in advanced writing by L1-Swedish English learners and native English speakers of English and found that, in terms of diversity, native speakers of English use a larger number of lexical bundles than non-native speakers of English in their writings. In another study, Qin (2014) investigated how advanced non-native English graduate students of applied linguistics at different levels of study use five-unit target lexical bundles in their academic papers and reported that "noun phrase with other post-modifier fragments" are used more frequently by writers at higher levels of study. Allan (2016) examined the three- and four-word lexical bundles found in graded readers, and investigated to what extent these bundles are affected by simplified language. He found that at B1 level, most lexical bundles are verb phrases than noun phrases. At B2 level and in FIC, however, noun phrase lexical bundles are predominant. He concluded that simplifying texts may influence the structural composition of four-word lexical bundles. That is, a higher level of simplification leads to more verb phrase lexical bundles in texts which is a feature of spoken language (Biber et al., 2004). Yet, in another study, Pérez-Llantada (2014) investigated the use of lexical bundles in expert academic writing for L1 English, L2 English, and L1 Spanish learners. Structurally, most lexical bundles were a combination of two structural units, where the last word of one unit is the beginning of the second structure (e.g., a function of the, the rest of the). Additionally, irrespective of the language variable, she found that all the lexical bundles follow the norms of academic written register. That is, the majority of the bundles comprised phrases, rather than clauses.

In terms of discipline, research shows that the use of lexical bundles varies across different disciplines. In other words, the use of lexical bundles is discipline specific (e.g., Hong & Hua, 2018; Hyland, 2008a). Investigating a corpus of 3.5 million words of doctoral dissertations, master's theses, and research articles from the four disciplines of electrical engineering, microbiology, business, and applied linguistics, Hyland (2008a) found that "writers in different fields draw on different resources to develop their arguments, establish their credibility and persuade their readers, with less than half of the top 50 bundles in each list occurring in any other list" (2008a, p. 20). In two other studies, Cortes (2002, 2004) compared research articles in soft and hard fields and found that research articles in hard fields feature much more lexical bundles than research articles in soft fields and with different structural and functional features.

In terms of genre, research shows that the use of lexical bundles varies across different genres. For example, Biber et al. (1999) found that the number of bundles in classroom instruction is twice as much as and four times bigger than the number of bundles in textbooks and academic prose, respectively. In another study comparing three different genres, i.e. theses, dissertations, research articles, Hyland (2008a) found that the bundles used in theses and dissertations are more phrasal, whereas the bundles used in research articles are more clausal.

The plethora of studies on lexical bundles altogether suggests that investigating lexical bundles is a very important area of inquiry. Yet, due to the inconsistencies found in these studies, reaching a comprehensive image of how lexical bundles are appropriately used requires more elaborate studies. To this end, this paper sought to explore the frequency and diversity of the use of lexical bundles in chemistry research articles from a structural perspective.

3. Methodology

3.1. Corpus

The study was based on chemistry research articles corpus (CRAC), developed specifically for this study with more than four million words. CRAC consists of published articles from Elsevier’s online platform ‘ScienceDirect’. The papers included in CRAC all come from Elsevier’s well-known journals. The papers are equally distributed across the four main subject areas of chemistry, i.e. analytical chemistry, organic chemistry, inorganic chemistry, and physical/theoretical chemistry. To develop the corpus, we selected 10 well-known journals from each subject area randomly, first. However, because of access issues, only eight journals in relation to analytical chemistry were accessible and were thus selected for inclusion in CRAC. Next, eight volumes from each periodical were picked out, except for analytical chemistry journals, for which 10 volumes from each journal were gathered. Consequently, 320 volumes, published from 2003 to 2009 were selected. Finally, the papers, published in one issue from each volume, were compiled in the corpus. Altogether, the number of research papers, compiled in CRAC, reached 1,185. The number of authors was not an issue in compiling CRAC. Hence, papers with a single contributor or many authors were all gathered in CRAC. Additionally, intercultural rhetoric was not considered as an issue in selecting the papers, given the fact that we assumed the members of a specific discipline or discourse community follow their group conventions. Moreover, the English language proficiency of the authors and the status of English as a first, second, or foreign language for them were not considered as well, as we assumed the ability to publish in English well-known journal common among all the contributors.

3.2. Analytical Procedure

A corpus-driven discourse analytic approach, compliant with Hyland’s (2008b) approach, was followed in the present study. The study was corpus-driven, due to the fact that identification of the bundles was not based on an established list of lexical bundles from prior research. Instead, bundles were identified, using corpus linguistic tools. On the other hand, the discourse analytic nature of the study is rooted in the structural analysis of lexical bundles in their specific contexts. Hence, a corpus-driven discourse analytic approach was followed to extract lexical bundles from CRAC and investigate their structural features. WordSmith (Scott, 2015) was used to generate lexical bundles and concordance lines from the corpus. Three criteria were considered for identifying bundles: length, frequency, and dispersion. Because four-word lexical bundles “offer a wider variety of structures and functions to analyze” (Hyland, 2012, p. 151), the length of lexical bundles was set at four. In previous studies, frequency ranges from 10 to 40 times per million words (e.g., Biber et al., 1999; Biber, et al., 2004; Cortes, 2004; Hyland, 2008a, b). Here, a minimum frequency of 20 times per million words was set. In previous studies, dispersion or range has been set from three to five (e.g., (e.g., Biber et al.,1999; Cortes, 2013). Here, occurrence in at least five different texts was set as the range cut-off. Hence, using WordSmith, we generated a list of four-word lexical bundles that occurred at least 20 times per million words in at least five different papers. Next, in a discourse analytic approach, we used WordSmith Concord to code the structural features of the generated lexical bundles, following Biber et al.'s (1999) taxonomy. According to Biber et al., a distinction is made between clausal and phrasal bundles. Phrasal bundles fall under one of the structural groups of 'noun phrase + *of*', 'other noun phrases', 'prepositional phrase + *of*', 'other prepositional phrases'. Clausal bundles, on the other hand, have the structures of 'passive + prepositional phrase fragment', 'anticipatory *it* + verb\adjective', '*be* + noun\adjectival phrase', 'adverbial clause', '*that* clause fragment', 'pronoun/noun phrase + *be*', or 'others'. The distribution of each pattern was then computed and compared with others. To ensure precision in coding the structural patterns of lexical bundles, we coded them independently. A Cohen's kappa of 0.92 was computed for inter-coder reliability. In cases where disagreements ensued in coding the bundles, a third coder's opinion was sought. Moreover, a third coder was invited to code a random selection of 30% of the bundles. Here, a Cohen's kappa of 0.90 was reached in terms of inter-coder reliability.

4. Results

A total of 223 four-word lexical bundles, totaling 37,756 individual cases, occurring at least 20 times per million words in at least five different articles were found in the four-million-word CRAC. More than 28 bundles occurred over 240 times per million words in CRAC, which is much higher than the set minimum frequency in this article. In addition, more than 86% of the bundles appeared in more than 50 different RAs. Among them, *in the presence of, in the case of, as a function of, on the other hand, as shown in figure, the reaction mixture was, are shown in figure, is shown in figure, on the basis of, was found to be, with respect to the, in the range of, as well as the, to a solution of, in the absence of, the formation of the, the presence of the, to the formation of, at room temperature for*, and *a function of the* were the top 20 lexical bundles, occurring at least more than 280 times in the entire corpus. Their frequency from the first to the 20th most frequent lexical bundle ranged from 1315 to 283. These bundles appeared in at least 150 different RAs. Table 1 shows the results of structural analysis of lexical bundles in CRAC.

Table 1: Structural Analysis of Lexical Bundles in CRAC

Structures	No. of bundles	Overall Freq.	Percentage (%)
Noun phrase + *of*	52	7242	19.18
Other noun phrases	8	1042	2.60
Prepositional phrase + *of*	29	8021	21.24
Other prepositional phrases	26	4352	11.52
Passive + prepositional phrase fragment	52	7916	20.86
Anticipatory *it* + verb\adjective	11	1550	4.35
Be + noun\adjectival phrase	8	1191	3.15
Adverbial clause	4	1080	2.86
That clause fragment	6	744	1.97
Pronoun/noun phrase + be	3	858	2.27
Others	24	3764	9.96
Total	223	37756	100

As Table 1 shows, more than 55% of all the bundles in the entire corpus were phrasal rather than clausal, i.e. noun phrases and prepositional phrases. The bundles consisting of 'prepositional phrases + *of*' were found to be used more frequently than other bundles in CRAC. Prepositional phrase bundles, in general, accounted for more than one-third of all the bundles in the corpus. The second position is occupied by other phrasal bundles, i.e. 'noun phrase + *of*' and 'other noun phrases', which accounted for 22% of all the bundles. It can also be seen from Table 1 that 'noun phrase + *of*' and 'prepositional phrase + *of*' were almost three times more common than 'other noun phrases' and 'other prepositional phrases'.

Clausal bundles constituted 35.21% of all the bundles in the corpus. As Table 1 shows, in terms of both frequency and diversity, bundles with 'passive + prepositional phrase' structure were the predominant structural category among all clausal lexical bundles, i.e. 'adverbial clause fragment', 'anticipatory *it* + verb/adjective phrase', '*that*-clause fragment', 'copula be + noun/adjective phrase', and 'pronoun/noun phrase + be'. Bundles with 'passive + prepositional phrase' structure constituted almost 60% of all the clausal bundles. All the structural groups of bundles are presented in more details below.

4.1. 'Prepositional Phrase + of' Lexical Bundles

Table 2 presents the frequent lexical bundles in chemistry RAs with 'prepositional phrase + *of*' structure.

Table 2: 'Prepositional Phrase + of' Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
in the presence of	1315	404
in the range of	355	216
in the case of	1100	424
as a function of	883	321
on the basis of	377	217
to a solution of	322	110
in the absence of	316	156
to the formation of	306	199
for the preparation of	262	126
for the determination of	217	103
to the presence of	216	162
on the surface of	1961	107
as a result of	195	140
for the formation of	193	105
for the synthesis of	155	89
in the form of	144	101
in terms of the	137	96
at the end of	136	94
by the presence of	126	100
to that of the	123	102
by the addition of	120	85
in the region of	100	57
to the loss of	88	52
by the reaction of	87	56
at a rate of	80	60
in the spectra of	103	50
with the exception of	85	63
with the increase of	128	72
in the formation of	156	108

Altogether, as Table 2 shows, there were 29 lexical bundles of this types in CRAC which accounted for about 21.24% of bundles in the entire corpus. These bundles took the first place in the corpus with an overall frequency count of 8021, and accounted for more than 38% of all the phrasal bundles. Some of the bundles in this group were extremely frequent in the corpus and were the most frequent bundles in the whole corpus. For example, *in the presence of*, *in the case of*, and *as a function of* were the first three most frequent bundles in CRAC. Most of the bundles in this structural group were extended to 5-word bundles. For example, *in the presence of* was usually part of a larger prepositional phrase such as *in the presence of the* and *in the presence of a*.

4.2. 'Other Prepositional Phrase' Lexical Bundles

Table 3 presents the frequency of 'other prepositional phrase' bundles in chemistry RAs.

Table 3: 'Other Prepositional Phrase' Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
on the other hand	720	416

with respect to the	368	226
at room temperature for	289	147
in agreement with the	217	165
at the same time	188	138
in the present work	183	111
in the gas phase	181	65
in good agreement with	177	126
in the present study	164	116
similar to that of	144	111
as a white solid	139	32
due to the presence	139	115
in the solid state	131	82
to the fact that	129	106
in order to obtain	129	109
in addition to the	127	104
in this case the	118	104
for the first time	105	71
as a result the	94	78
in contrast to the	94	80
in accordance with the	91	79
in comparison with the	87	69
to a stirred solution	87	36
in order to determine	82	73
due to the fact	81	71
as well as in	84	68

Overall, as Table 3 shows, there were 26 lexical bundles of this type in the corpus. They comprised 11.52% of all the bundles in CRAC. The variety and overall frequency of these bundles were less than the previous bundles, i.e. ‘prepositional phrase + *of*’. Some of the bundles in this group were very frequent in the corpus. These include *on the other hand*, *with respect to the*, and *at room temperature for*. For example, *on the other hand* with an overall frequency of 720 times took the fourth position among the first 10 most frequent bundles. Altogether, bundles with prepositional phrases accounted for more than 34% all the bundles in the corpus.

4.3. ‘Noun Phrase + of’ Lexical Bundles

Among phrasal bundles, the second most frequent category was characterized by the structure ‘noun phrase + *of*’. Table 4 presents the frequency and diversity of different lexical bundles with ‘noun phrase + *of*’ structure.

Table 4: ‘Noun Phrase + of’ Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
the formation of the	312	168
the presence of the	312	210
a function of the	283	153
the effect of the	272	169
the case of the	257	156
the presence of a	246	166
the surface of the	236	136
the formation of a	208	139
the structure of the	200	148
the influence of the	187	119
the nature of the	186	131
the results of the	158	121
the basis of the	157	104
the stability of the	157	99
a wide range of	155	117
one of the most	148	130

the increase of the	142	90
the concentration of the	141	102
the intensity of the	136	89
the crystal structure of	125	76
the values of the	125	85
the size of the	124	82
the ratio of the	123	95
the determination of the	118	92
the temperature of the	112	75
the value of the	112	86
a result of the	108	83
the analysis of the	105	80
a flow rate of	104	78
the position of the	102	79
a small amount of	98	84
the shape of the	97	66
the temperature dependence of	95	55
the use of a	94	75
the dependence of the	93	68
a stirred solution of	92	39
the composition of the	92	74
the thickness of the	92	51
a large number of	91	77
a function of time	90	57
the length of the	90	60
a heating rate of	89	57
the ir spectra of	88	54
the slope of the	87	72
the spectra of the	87	54
the decrease of the	86	65
the ph of the	80	53
the sum of the	80	60
the stability of the	157	99
temperature dependence of the	97	51
the end of the	131	91
a function of temperature	85	47

Bundles with ‘noun phrase + *of*’ structure constituted 19.18% of all the bundles in the corpus. In terms of variety, they outnumbered all the other phrasal bundles. As Table 4 shows, there were 52 different types of these bundles in the corpus which is much higher than the 32 different bundles with ‘prepositional phrase + *of*’ structure. Yet, the overall frequency of these bundles was less than ‘prepositional phrase + *of*’ bundles. Some of the most frequent bundles of this group were *the formation of the*, *the presence of the*, *a function of the*, and *the effect of the.* All the bundles of this type were preceded by articles *the* and *a* and followed by *the, of*, and *a.*

4.4. ‘Other Noun Phrase’ Lexical Bundles

Table 5 presents the frequency and diversity of ‘other noun phrase’ lexical bundles in CRAC.

Table 5: ‘Other Noun Phrase’ Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
the fact that the	212	170
an increase in the	178	113
a decrease in the	131	91
good agreement with the	120	93
an important role in	109	92
the difference between the	108	83
the increase in the	94	68

chromatography on silical gel	90	30

As Table 5 shows, compared to other phrasal bundles, i.e. 'noun phrase + *of*', 'prepositional phrase + *of*', and 'other prepositional phrases', this group of bundles were the least varied and frequently used ones in the corpus. There were 8 types of lexical bundles with this structure in CRAC. With a total frequency count of 1042, these bundles accounted for only 2.60% of all the bundles in the whole corpus.

4.5. 'Passive + Prepositional Phrase Fragments' Lexical Bundles

Table 6 shows the frequency and diversity of bundles with 'passive + prepositional phrase fragments' structure in CRAC.

Table 6: 'Passive + Prepositional Phrase Fragments' Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
are shown in fig	513	327
is shown in fig	501	315
was found to be	370	214
are given in table	261	177
were recorded on a	249	177
are listed in table	227	155
was added to the	207	141
be attributed to the	191	142
were carried out in	185	153
can be attributed to	174	127
stirred at room temperature	168	125
can be used to	166	79
are summarized in table	164	127
are shown in table	162	125
were found to be	160	123
can be seen in	159	121
was used as the	153	103
used without further purification	151	119
was added to a	146	137
used in this study	135	71
was stirred at room	132	93
was used as a	130	63
was used for the	126	110
were carried out using	120	98
reported in the literature	117	109
is based on the	115	84
be explained by the	114	101
can be explained by	114	91
can be seen from	112	80
were carried out at	112	99
is related to the	99	80
was carried out on	99	58
added to a solution	98	47
was observed in the	98	77
were performed on a	98	86
are presented in table	97	72
be seen in fig	97	70
were used as received	93	88
carried out in a	90	79
are presented in fig	89	65
can be seen in fig	89	62
were carried out with	89	83
was purified by flash	88	21

was carried out using	87	76
was carried out with	87	66
was added to a solution	85	41
were carried out on	85	77
be related to the	81	69
is found to be	81	57
was found to be	370	214
was carried out in	102	88
solved by direct methods	80	67

As Table 6 shows, totally, there were 52 different bundles with the structure ‘passive verb followed by a prepositional phrase fragment’ in the corpus. These clausal bundles accounted for 20.86% of all the bundles in CRAC. After ‘prepositional phrase + *of*’ bundles, they were the most frequent lexical bundles in CRAC. These bundles were much more frequent than other clausal bundles, i.e. ‘adverbial clause fragment’, ‘anticipatory *it* + verb/adjective phrase’, ‘*that*-clause fragment’, ‘copula Be + noun/adjective phrase’, ‘pronoun/noun phrase + Be’. They were more than 10 times as common as bundles with ‘*that*-clause fragment’ structure. Some of the bundles in this category, such as *are shown in figure*, *is shown in figure*, and *was found to be* were very frequent and appeared in many different texts in the corpus, occurring 128, 125, and 92 times per million words, respectively, and in 327, 315, and 214 different texts, respectively.

4.6. ‘Anticipatory it + Verb\Adjective’ Lexical Bundles

Table 7 shows the frequency and diversity of ‘anticipatory *it*’ bundles in CRAC.

Table 7: ‘Anticipatory it’ Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
it can be seen	267	155
it was found that	246	168
it is possible to	172	126
it should be noted	142	105
it is important to	130	100
it is well known	120	105
it is clear that	114	92
it is known that	106	91
it has been shown	90	75
it was observed that	82	54
it can be concluded	81	63

As Table 7 shows, there were 11 different ‘anticipatory *it*’ bundles in the corpus. They formed 4.35% of all the bundles. After ‘passive verb + prepositional phrase’, this group of bundles were the most used clausal bundles, both in terms of frequency and variety. Some of the most frequent bundles of this group were *it can be seen*, *it was found that, and it should be noted.* Most of these bundles were extended into 5- and 6-word bundles. For example, *it can be seen* and *it should be noted* are part of the larger bundles *it can be seen that* and i*t should be noted that*, respectively.

4.7. ‘That-Clause Fragment’ Lexical Bundles

Table 8 shows the frequency and diversity of ‘that-clause fragment’ bundles in CRAC.

Table 8: ‘That-Clause Fragment’ Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
can be seen that	178	110
should be noted that	135	101
was found that the	133	101
is well known that	111	96
be seen that the	106	79
be noted that the	81	69

As Table 8 shows, there were only six bundles of this type in the whole corpus. They constituted less than 2% of all the bundles in CRAC. The most frequent bundle in this group was *can be seen that*, with a total frequency of 178 in the whole corpus. In most cases, this bundle was part of a larger bundle *can be seen that the.*

4.8. 'Be + Noun\Adjectival Phrase' Lexical Bundles

Table 9 shows the frequency and diversity of 'Be + noun\adjectival phrase' lexical bundles in CRAC.

Table 9: 'Be + Noun\Adjectival Phrase' Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
is due to the	222	161
be due to the	192	148
is one of the	168	142
is consistent with the	157	122
may be due to	147	104
is similar to that	80	66
is in agreement with	143	123
are in good agreement	85	72

As Table 9 shows, there were only eight different types of bundles with 'Be + noun\adjectival phrase' structure in the corpus. Overall, they occurred 1191 times and accounted for 3.15% of all the bundles in CRAC. As Table 9 shows, *is due to the* was the most commonly used bundle of this type, occurring 222 times in the whole corpus.

4.9. 'Adverbial Clause Fragment' Lexical Bundles

Table 10 shows the frequency and diversity of lexical bundles with adverbial clause fragments in the corpus.

Table 10: 'Adverbial Clause Fragment' Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
as shown in fig	646	321
as can be seen	213	121
as shown in scheme	82	59
as shown in table	139	108

As Table 10 shows, there were only four different bundles with adverbial clause fragments in the corpus. Yet, in terms of occurrence, they were much more frequent (1080 times) than the six bundles with 'that-clause' structure (744 times). Although these bundles constituted only 2.86% of all the bundles, some of them were extremely frequent in the whole corpus. As Table 10 shows, the most frequent bundle in this group, i.e. *as shown in figure,* was the fourth most frequent bundle in the entire corpus. It occurred about 646 times in CRAC and in 321 different texts.

4.10. 'Pronoun/Noun Phrase + Be' Lexical Bundles

Table 11 shows the frequency and diversity of lexical bundles with 'pronoun/noun phrase + Be' structure in the corpus.

Table 11: 'Pronoun/Noun Phrase + Be' Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
the reaction mixture was	572	173
the organic layer was	180	52
the crude product was	106	37

As Table 11 shows, there were only three groups of lexical bundles with ‘pronoun/noun phrase + Be’ structure. These bundles had the least variety among all the structural groups and accounted for 2.26% of all the bundles in the corpus. *The reaction mixture was* is the most frequent bundle in this group with 572 occurrences in the whole corpus.

4.11. ‘Other’ Lexical Bundles

Table 12 shows the frequency and diversity of lexical bundles in CRAC that could not be grouped into phrasal or clausal categories. These bundles are referred to as ‘others’ in Biber et al.’s (1999) taxonomy.

Table 12: ‘Others’ Lexical Bundles in CRAC

Lexical bundles	Freq.	No. of texts
as well as the	344	239
the other hand the	253	196
spectra were recorded on	233	176
the mixture was stirred	226	115
mixture was stirred for	197	102
mixture was stirred at	191	78
in this paper we	145	134
reaction mixture was stirred	183	75
than that of the	179	117
nmr spectra were recorded	169	153
was added and the	161	69
and the mixture was	148	85
higher than that of	145	109
at room temperature and	138	108
experiments were carried out	138	119
measurements were carried out	130	102
mmol was added to	119	53
ml the combined organic	113	30
the solvent was removed	109	46
and the reaction mixture	100	43
ml was added to	87	53
probably due to the	87	77
in this work we	86	75
lower than that of	83	66

As Table 12 shows, there were 24 lexical bundles in the corpus that did not fell under phrasal or clausal bundles. Totally, they accounted for 9.58% of all the bundles in the corpus. Some of these bundles were very frequent in chemistry RAs. For example, *as well as the* and *the other hand the* appeared 344 and 253 times in the whole corpus, respectively.

6. Discussion and Conclusion

The present corpus-driven study explored the structural features of lexical bundles, used in chemistry research articles. Altogether, the results revealed 223 frequent lexical bundles in the four-million-word corpus of chemistry RAs, totaling 37,756 tokens. More than 55% of the bundles were phrasal. On the other hand, 35% of the bundles were clausal. This is in keeping with the results of Pan et al. (2016), Qin (2014), Cortes (2013), and Zare and Naseri (2020). As Biber et al. (2004) note, the existence of a large number of verb phrases is a feature of spoken language, rather than written texts. Moreover, “careful integration of information in academic prose requires the use of noun phrases and prepositional phrases, which leads to a shift from clausal style to phrasal style in academic prose” (Pan et al., 2016, p. 65). This finding mirrors the results of previous studies that pointed to the dominance of phrasal lexical bundles in academic texts (e.g., Biber & Conrad, 1999; Biber et al., 1999, 2004; Esfandiari & Barbary, 2017; Pan et al., 2016; Qin, 2014).

Among the bundles, in the presence of, in the case of, and as a function of were the first three most frequent formulaic sequences in chemistry RAs. All these bundles took the pattern of ‘prepositional phrases + of’ which was used more frequently than other structural categories in CRAC, accounting for more than 38% of all the phrasal bundles. This is in keeping with the results of Zare and Naseri (2020) and Hyland (2008a, b). Bundles with the structure ‘passive verb followed by a prepositional phrase fragment’ were the second most frequent formulaic sequences in the entire corpus, constituting more than 20% of all the bundles in CRAC. This mirrors the results of Biber et al. (1999). In line with Biber et al., the results of the present study show that verb phrases mostly comprise passive constructions, followed by prepositional phrases. This is opposite to what Zare and Naseri (2020) found. The corpus Zare and Naseri used mainly constituted articles from soft disciplines, i.e. linguistics and applied linguistics, whereas the articles in CRAC were compiled from hard disciplines, i.e. chemistry. Such bundles are mostly used to refer to graphical or tabular data in the articles of hard disciplines (Hyland, 2008a). The third and second most frequent category in CRAC and among phrasal bundles, respectively, was characterized by bundles with the structure ‘noun phrase + of’. These bundles which constituted more than 19% of all the bundles in the corpus outnumbered all the other phrasal bundles in variety. These results mirror the findings of Biber et al. (1999), Biber (2010), Esfandiari and Barbary (2017), Jalali and Zarei (2016), Ädel and Erman (2012), and Zare and Naseri (2020). Ädel and Erman found most of four-word lexical bundles in academic writing to incorporate noun or prepositional phrases. As Biber (2010) notes, “70% of the common bundles in academic prose consist of a noun phrase with an embedded prepositional phrase fragment (e.g., the nature of the) or a sequence that bridges across two prepositional phrases (e.g., as a result of)” (p. 172). Moreover, the bundle on the other hand was also among the first 10 most frequent bundles in the whole corpus. This is in keeping with the findings of Hyland (2008a) who found this bundle as the most frequent formulaic sequence in his corpus of electrical engineering, microbiology, business studies, and applied linguistics written texts. The predominance of noun and prepositional phrases in this study is in line with Qin’s (2014) observation that “noun phrases with post-modifier fragments, including prepositional phrases or past participle phrases, are less likely to appear in non-native graduate writers’ writing than in expert writers’ academic discourse” (p. 225). Qin relates this to “the inherently complex structural forms of these bundles, which require writers to pack their message or information in the most economical manner, an important feature of academic writing” (Biber et al., 1999, as cited in Qin, 2014, p. 225).

Among clausal bundles, those with the structure ‘passive + prepositional phrase’ were the predominant structural category in terms of both frequency and diversity. Some bundles with adverbial clause fragments such as as shown in figure, though accounting for less than 3% of all the bundles, were among the most frequent bundles in the whole corpus. Additionally, lexical bundles with ‘that clause fragments’ were the least frequently used formulaic sequences in chemistry RAs.

In terms of diversity, ‘noun phrase + of’ and ‘passive + prepositional phrase fragment’ bundles were the most varied and bundles with ‘pronoun/noun phrase + be’ structure were the least varied formulaic sequences in the corpus. What is important to note is that in terms of frequency, lexical bundles with prepositional phrases outnumbered bundles with noun phrases. However, in terms of diversity, lexical bundles with noun phrases outnumbered bundles with prepositional phrases. This may be taken to indicate that frequency and diversity correlate with the type of phrasal lexical bundle used. Altogether, the results of this research, along with other studies, suggest that different discourses are associated with different sets of lexical bundles with different frequency and diversity, due to the different communicative functions they follow (Tseng, 2018).

Writing a well-developed research paper in English is a very important, yet demanding task. On the other hand, awareness of the recurrent phrases, i.e. lexical bundles, used in research papers, is of great help to the writers of this discourse. Yet, developing a corpus-driven list of lexical bundles is not an easy task. Hence, the findings of this corpus-driven analysis, though by no means conclusive, are useful on many levels. First, the list of generated lexical bundles can be used as basis for comparative research on lexical bundles in other genres. Second, the findings can be used in EAP materials development and pedagogy in chemistry. Therefore, EAP material developers may use the generated list of bundles in their materials for chemistry students and EAP instructors may focus their chemistry students' attention on these bundles when teaching how to write an effective research paper in English. Practice with the frequent lexical bundles and their contextualized examples raises the chemistry students' awareness of the kind of language they need to develop in order to be able to publish the results of their research studies.

The findings of this study need to be treated with some caution, due to the following limitations and delimitations. First and foremost, the corpus on which we based our analysis was limited to journals, published by Elsevier only. Second, identifying lexical bundles in this study was only based on their length, frequency, and dispersion. Calculating MI score was not possible. Computing the MI score helps us understand if the words that occur together in a phrase occur more often than expected by chance (Ellis, Simpson-Vlach, & Maynard, 2008, p. 380). Third, due to the subjective nature of coding, different measures need to be taken in order to ensure precision. Fourth, other aspects such as intercultural rhetoric, the number of authors per article, their English language proficiency level, and the status of English as their first, second, or foreign language were not considered in the paper. Hence, future studies need to base their analyses on more comprehensive corpora, compiled from articles, published by well-known publishers, compute the MI score, take different measures to increase the objectivity of coding, and consider aspects of the contributors of papers such as intercultural rhetoric, the number of authors per article, their English language proficiency level, and the status of English as their first, second, or foreign language. Further research may also investigate the diversity of lexical bundles across the sub-fields of chemistry and other disciplines.

References

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes, 31*, 81-92.

Allan, R. (2016). Lexical bundles in graded readers: To what extent does language restriction affect lexical patterning? *System, 59*, 61-72.

Biber, D. (2010). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine & H. Narrog (eds.), *The Oxford handbook of linguistic analysis* (pp. 159-191). Oxford: Oxford University Press.

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*, 263-286.

Biber, D., & Conrad, S. (1999). Lexical bundles in conversation and academic prose. In H. Hasselard & S. Oksefjell (eds.), *Out of corpora: Studies in honour of Stig Johansson* (pp. 181-189). Amsterdam: Rodopi.

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A. Wilson, P. Rayson, & T. McEnery (eds.), *Corpus linguistics by the lune: A festschrift for Geoffrey Leech* (pp. 71-92). Frankfurt, Germany: Peter Lang.

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45-61.

Cortes, V. (2002). *Lexical bundles in academic writing in history and biology* (Doctoral dissertation). Flagstaff, Arizona: Northern Arizona University.

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*, 397-423.

Cortes, V. (2013). The purpose of this study is to: connecting lexical bundles to moves in research article introductions. *Journal of English for Academic Purposes, 12,* 33-43.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in second language acquisition, 18,* 91-126.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24*, 143-148.

Ellis, N. C., & Simpson-Vlach, R. (2008). Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*, 375-396.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375–396.

Esfandiari, R., & Barbary, F. (2017). A contrastive corpus-driven study of lexical bundles between English writers and Persian writers in psychology research articles. *Journal of English for Academic Purposes, 29*, 21-42.

Esfandiari, R., & Moein, G. (2016). A corpus-driven investigation into lexical bundles across research articles in food science and technology. *English Language Teaching, 3*(1), 1-30.

Hong, J. (2019). Structural use of lexical bundles in the rhetorical moves of L1 and L2 academic writing. *English Teaching, 74*(3), 29-54.

Hong, A. L., & Hua, T. K. (2018). Specificity in English for academic purposes (EAP): A corpus analysis of lexical bundles in academic writing. *3L: The Southeast Asian Journal of English Language Studies*, *24*(2), 82-94.

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

Hyland, K. (2008b). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18,* 41-62.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics, 32*, 150-169.

Jablonkai, R. (2010). English in the context of European integration: A corpus-driven analysis of lexical bundles in English EU documents. *English for Specific Purposes, 29,* 253-267.

Jalali, Z., & Moini, M. R. (2018). Structure of lexical Bundles in introduction section of medical research articles. *Procedia Social and Behavioral Sciences, 98*, 719-726.

Jalali, Z., Moini, M. R., & Alaee Arani, M. (2015). Structural and functional analysis of lexical bundles in medical research articles: A corpus-based study. *International Journal of Information Science and Management, 13*(1), 51-69.

Jalali, H. & Zarei, G. (2016). Academic writing revisited: a phraseological analysis of applied linguistics high-stake genres from the perspective of lexical bundles. *The Journal of Teaching Language Skills (JTLS), 7*(4), 87-114.

Li, J., & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: a longitudinal case study. *Journal of Second Language Writing, 18,* 85-102.

Pan, F., Reppen, A., & 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.

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes, 14*, 84-94.

Qin, J. (2014). Use of formulaic bundles by non-native English graduate writers and published authors in applied linguistics. *System, 42,* 220-231.

Rahimi Azad, H., & Modarres Khiabani, S. (2018). Lexical bundles in English abstracts of research articles written by Iranian scholars: Examples from humanities. *Iranian Journal of Applied Language Studies, 10*(2), 149-174.

Ruan, Z. (2016). Lexical bundles in Chinese undergraduate academic writing at an English medium university. *RELC Journal, 48(3),* 327-340.

Scott, M. (2015). WordSmith tools version 6. Lexical Analysis Software.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12,* 214-225.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language, 10*, 65-108.

Tseng, M. (2018). Creating a theoretical framework: On the move structure of theoretical framework sections in research articles related to language and linguistics. *Journal of English for Academic Purposes, 33*, 82-99.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zare, J., & Naseri, Z. S. (2020). Lexical bundles in English review articles. *Iranian Journal of English for Academic Purposes, 9*(1), 41-56.