Iranian Journal of Applied Language Studies
Vol 10, No 2, 2018, pp. 149-174
http://ijals.usb.ac.ir

Lexical Bundles in English Abstracts of Research Articles Written by Iranian Scholars: Examples from Humanities

Hossein Rahimi Azad
PhD. Candidate,
Science and Research Branch,
Islamic Azad University, Tehran, Iran
hossein.rahimiazad@srbiau.ac.ir

Shahram Modarres Khiabani
Corresponding author, Assistant Professor,
Karaj Branch, Islamic Azad University, Karaj, Iran
shahram.modarres@kiau.ac.ir

Abstract

This paper investigates a special type of recurrent expressions, lexical bundles, defined as a sequence of three or more words that co-occur frequently in a particular register (Biber et al., 1999). Considering the importance of this group of multi-word sequences in academic prose, this study explores the forms and syntactic structures of three- and four-word bundles in English abstracts written by Iranian scholars in three disciplines; business studies, history, and linguistics. Applying a frequency-based approach suggested by Biber et al. (1999), the data were drawn from a corpus consisting of 660 English abstracts of research articles published by various leading academic journals. As regards syntactic structure, the structural characteristics of the lexical bundles were first explored through careful analysis and then classified using a modified version of Biber et al.'s (1999) structural framework. The study also compares the forms, frequencies, and structures of the lexical bundles in the corpus to those found in Hyland's (2004) research article corpus. This comparison shows that not only are there relatively large differences between the two corpora in the forms and frequencies of bundles employed, but also there are some discrepancies in terms of distribution of the different structural types and tokens.

Keywords: Lexical Bundles, Formulaic Expressions, Recurrent Expressions, Corpus Linguistic, Academic Writing

Received: March 2017; Accepted: December 2017

1. Introduction

Language, whether written or spoken, is formulaic (Sinclair, 1991, p.108). Although there is uncertainty on the extent to which language is formulaic, the formulaic expression, more specifically, is an umbrella term covering a wide variety of word sequences which are stored and retrieved as whole chunks rather than being subject to productive and analytical processing (Wray & Perkins, 2000, p.2). The study of formulaic expressions has a long history in applied linguistics, dating back to Firth (1951), who popularized the term 'collocation' along with the famous slogan that 'you shall know a word by the company it keeps' (Firth, 1951, p.11).

In general, lexical patterning is believed to play a key role in fluent linguistic production in a given genre (Haswell, 1991, p. 236). According to Hyland (2008), the recurrent use of formulaic expressions and pre-fabricated sequences is a way of facilitating communicative competence by making language more predictable and reducing processing time (Hyland, 2008, p. 5). On the other hand, as Haswell (1991, p. 236) believes, 'there can be little doubt that as writers mature they rely more and more on formulaic expressions and collocations'. Thus, from a pedagogic perspective, it seems reasonable to argue that learning the preferred ways of combining words in a given register or genre can help learners make appropriate linguistic choices and gain a high level of native-like proficiency in language learning.

The present study intends to investigate a particular type of formulaic expressions, lexical bundles, defined as a sequence of three or more words like *the nature of*, *as a result of*, and *in the presence of* that co-occur frequently in a particular register, which are not idiomatic nor complete structural units (Biber et al., 1999, pp. 988-991). Another important feature of lexical bundles is that they vary across genres. For instance, Biber et al. (1991, pp. 988-991) finds that noun and

prepositional phrases such as *the nature of*, and *in the presence of* are particularly salient in academic prose but not in conversation. Frequency of occurrence is the defining characteristic for lexical bundles; in order for a word sequence to count as a bundle, it must occur at least twenty times in a million words with the additional requirement that this rate of occurrence be realized in at least five different texts to guard against idiosyncratic or repetitive uses (Cortes, 2004, pp. 399-400). Also, fixed in form is among other properties of bundles. But as Cortes (Ibid) points out, this fixedness is a result of the frequency criteria applied during the bundle extraction process and is thus different from the fixedness characterizing other word combinations.

Here, the focus is on the use of lexical bundles in English academic discourse and, more specifically, in abstracts of research papers. Considering that English abstracts are of special interest for their relevance in most academic texts, even where English is not the official language, this study intends to explore the forms and syntactic structures of three- and four-word bundles in English abstracts written by Iranian scholars in three disciplines; business studies, history, and linguistics. Applying a frequency-based approach suggested by Biber et al. (1999) and using a modified version of Biber et al.'s structural framework (1999), the study also compares the forms, frequencies, and structures of the lexical bundles in the *research corpus* to those found in Hyland's (2004) research article corpus.

This study, therefore, is seeking to address the following questions:

1. To what extent is there evidence to support similarity or dissimilarity in the forms, and structures of bundles used in the *research corpus* and of those found in Hyland's (2004) research article corpus?
2. How can bundles used in the *research corpus* be classified according to their structural features?

Following this first introductory section, section two presents a brief overview of previous research on lexical bundles. Section three is devoted to introducing the theoretical framework, Biber et al.'s (1999) approach. Then, in section four, we describe the methodology used to identify lexical bundles based on a frequency-based approach suggested by Biber et al. (1999). Section five provides results and conclusions drawn from comparing the forms, frequencies, and structures of the lexical bundles in the *research corpus* to those found in Hyland's (2004) research article corpus. Finally, section six is concluding remarks.

2. Literature Review

Lexical bundles were first defined and described in detail by Biber et al. (1999) in their exhaustive corpus-based study of English grammar. Considering that it was their first time to identify lexical bundles so they were more conservative and only adopted a minimum frequency of ten times per million words and an occurrence in at least five different texts in each register. As well as structural grouping of bundles, Biber et al. (1999) also compared their uses and distributions across a wide range of registers like conversation, fiction, news, academic prose, and non-conversational speech. With regard to their structure, these authors proposed a structural classification for lexical bundles based on their strong grammatical correlates, which will be mentioned in detail in next section.

Since 1999, some attempts have also been made to explore possible differences and similarities in the use of bundles between different disciplinary fields (soft and hard ones), registers (spoken and written), genres, and different degrees of writing expertise. Among studies focusing on disciplinary variations in the use of bundles, Cortes (2002, 2004) found that research articles in biology as one of the hard fields employed bundles much more than those of history, which is a soft field. Her study

also showed some major structural and some few functional differences between these two disciplines in the uses of bundles.

In the studies of variations across registers, Biber et al. (1999) compared conversation and academic prose, while Biber et al. (2004) worked on two other registers; classroom teaching and textbooks. These two studies together indicated that the number of lexical bundles in classroom teaching was almost twice more than that of conversation and around four times more than that of textbooks and academic prose.

With regard to possible generic variations in the use of bundles, Hyland (2008) investigated the frequency, forms and functions of lexical bundles in a large corpus composed of research articles, master theses, and doctoral dissertations in four different disciplines (i.e., electric engineering, microbiology, business studies, and applied linguistics). He then modified Biber et al.'s (2004) classification to create categories that better represented the lexical bundle functions he found in his corpus of research writing. Comparing these three different genres, Hyland (2008) showed that master theses employed bundles more than dissertations and much more than research articles. With regard to their structure, he realized that unlike research articles, bundles in student genres were more phrasal than clausal. His findings led him to question the notion of a core academic phrasal lexicon and call for a discipline-specific approach to the teaching of lexical bundles.

Cortes' study, 'Lexical bundles in published and student disciplinary writing: Examples from history and biology' is the sole study focusing specifically on examining possible variations in the use of bundles across different degrees of writing expertise. Cortes (2004) analyzed the forms and functions of the most frequent four-word bundles in published history and biology articles, which she called *target bundles,* and examined their uses in text written by students at three different levels in the same disciplines. Her findings showed that student rarely

used target bundles in their writing, and those that they used were employed in a different way than in professionally written texts.

These studies together indicate that the frequent and appropriate use of lexical bundles is an important component of fluent linguistic production in an academic environment, 'helping to shape meanings in specific contexts and contributing to our sense of coherence in a text' (Hyland, 2008, p.4).

The issue of lexical bundles has received considerable attention among Iranian researchers and scholars especially in recent years. For instance, Jalali (2009) carried out a study on lexical bundles in different genres of research articles, master dissertations, and doctoral theses on applied linguistics. Also, Valipoor (2010) identified lexical bundles in the genres of research articles in the discipline of chemistry. She found that bundles were associated with specific functions in sections of research articles and each section drew on specific set of bundles.

3. Theoretical Framework

As mentioned in pervious section, lexical bundles were first coined and explored by Biber and colleagues in a chapter of the *Longman Grammar of Spoken and Written English* (LGSW) (1999), where they define lexical bundles as 'bundles of words show a statistical tendency to co-occur' (1999, p. 989) and as 'recurrent expressions, regardless of their idiomaticity, and regardless of their structural status' (Ibid. 990). This definition, in fact, is based on a frequency-driven approach and the frequency cut-off they applied in their study to identify three- to six-word lexical bundles. Although frequency cut-offs are somewhat arbitrary and range between ten and forty instances per million words, the minimal cut-off set by Biber et al. (1999) was at least ten times per million words. Another condition Biber and colleagues used to identify lexical bundles is dispersion, meaning that a recurring lexical sequence must occur in multiple texts within a register to qualify as a lexical bundle (Salazar,

2011, p. 32). Specifically, according to Biber et al. (1999), in order for a recurring lexical sequence such as *on the other hand, in the case that* to count as a bundle, it must occur at least ten times in a corpus made of one million words with the additional requirement that this rate of occurrence be realized in at least five different texts in each register (Biber et al., 1999, p. 998-1024). In fact, Biber and colleagues (1999) identified frequently occurring lexical sequences in the conversation and academic sections of the *Longman Spoken and Written English Corpus* (LSWE), with each section containing around five million words using a computer program and they found that the longer the bundle, the lower was its frequency. Their findings showed that three-word lexical bundles occurred more than four-word lexical bundles in the conversation as well as in academic sections of the LSWE. They also reported that 'there were almost ten times as many three-word lexical bundles as four-word bundles, and about ten times as many four-word bundles as five-word bundles' (Ibid. 993). Concerning their structure, Biber and colleagues (1999) found that lexical bundles were, in most cases, not complete structural units and that most bundles bridge two structural units, that is, the last word of the bundle was often the first element of the following structure (Ibid. 993-1000). However, Biber et al. (1999) also realized that lexical bundles had strong structural correlates, which facilitated their grouping into several basic structural types. These authors proposed a structural classification for lexical bundles based on these typical grammatical correlates as seen in table 1. Based on Biber et al.'s structural framework (1999), in fact, lexical bundles are classified into two broader categories; phrasal and clausal bundles.

Table 3.1. *Structural Classification of Lexical Bundles in Academic Prose* (Biber et al., 1999, pp. 1015-1024)

	Structural patterns	Examples
1	Noun phrase with *of*-phrase fragment	*the end of the, the form of a*
2	Noun phrase with other post-modifier fragment	*the way in which, those of you who*
3	Prepositional phrase with embedded *of*-phrase fragment	*in the presence of, as a result of*
4	Other prepositional phrase (fragment)	*in the present study, on the other*
5	Passive verbs + prepositional phrase fragment	*is based on the, is shown in figure*
6	Copula *be* + [noun phrase/adjective phrase]	*is the same as, was no significance difference*
7	[Dummy there/pronoun/noun phrase] + be + noun phrase fragment	*there are a number, this is not the*
8	Anticipatory *it* + [verb phrase/adjective phrase] or Anticipatory it + [verb phrase/adjective phrase] + [to/that]	*It should be noted, it can be seen, it is clear that, it is possible to, it may be necessary, it should be noted that*
9	(Verb phrase +) *that*-clause fragment	*that there is a, that it is no*
10	Adverbial clause fragment	*as shown in figure, as we have seen*
11	Other expressions	*as well as the, the presence or absence*

As shown in table 3.1, there are bundles such as *the presence or absence* that 'do not fit neatly into any of the other categories' (Ibid. 1024), and thus, these bundles are called *other expressions* according to Biber et al.'s structural framework (1999).

Biber et al. (1999) also found that the grammatical correlates in lexical bundles differed considerably depending on the register. In other words, most of the bundles in conversation were clausal, of the type *pronoun + verb + complement*, such *I want you to*, and *it's going to be*, while in academic prose, most lexical bundles were phrasal, parts of noun phrases and prepositional phrases, as in the case of *on the basis of*, and *on the other hand* (Biber et al., 2004, p. 377).

In addition, the register comparisons carried out by Biber and colleagues (1999) had shown *th* at most of the bundles they identified in academic prose bridged two structural units, such as a noun phrase or beginning of a prepositional phrase. In other words, most lexical bundles in academic prose were found to consist of nominal or prepositional elements that co-occur in highly productive frames, such as *the….of the….* in which these two empty slots could be filled by many content words, as in the case of *the size of the*, and *the purpose of the* (Biber et al., 1999, pp. 991-993).

4. Methodology

4.1. Corpus Used In the Study

The data used in this study were drawn from two collections of academic writing. One corpus, which hereafter will be referred to as *research corpus*, comprises 660 English abstracts of research articles written by Iranian scholars, thus non-native speakers of English ,in three disciplines; business studies, history, and linguistics all of which were published by various leading academic journals between 2006 and 2011 (see table 4.1). In other words, all of those Iranian scholars as professional academic writers have a high degree of proficiency in academic English writing.

The other corpus used in this paper is Hyland's (2004) research article corpus, a 1.2 million-word corpus, composing of a collections of research articles written by native speakers of English in different disciplines (Simpson-Vlach & Ellis, 2010, p. 10). Since one of the research goals of this investigation is to identify Iranian scholars' deviant uses of lexical bundles in the English abstract they write for publication, a list of the most frequent lexical bundles in Hyland's (2004) research article corpus is considered as *reference bundles* for the purposes of comparison. Considering that there is a remarkable consistency between *research corpus* and

Hyland's (2004) research article corpus in terms of subject matter, register, and genre, it seems that we can make a meaningful comparison between the two corpora in the forms, and structures of bundles employed.

Table 4.1 *Composition of the Corpus of English Abstracts Written by Iranian Scholars* (Composition of *Research Corpus*)

Disciplines	Number of abstracts	Tokens
Business studies	220	46848
History	218	46897
Linguistics	222	46622
Total	660	149367

4.2. Identification of Lexical Bundles

Following previous corpus-based research on lexical bundles, this study focuses on three- and four-word lexical bundles. The main rational behind this decision is that many four-word bundles hold three-word bundles in their structures (Cortes 2004:401). In other words, three-word bundles are extremely common and usually incorporated into four-word sequences, and as Cortes (Ibid) points out, they present a wider variety of structures to analyze.

For the selection of lexical bundles, it is a common practice to set a frequency cut-off in the corpus and various values have been adopted in the literature. As mentioned before, Biber et al. (1999, pp. 988-1024) use the minimum frequency of ten occurrences in a corpus made of one million words in at least five different texts in each register. However, since he and colleagues (2004, p. 376) adopt a relatively high frequency cut-off of forty times per million words, we take the average of these two values for the present study, setting a minimum raw frequency of four instances in at least four different abstracts of the *research corpus*. Of course, since the two corpora used in this study are of distinctly different sizes, a normalization

procedure (a normalized frequency per 100,000 words) has been employed for comparing frequencies of individual bundles in the two corpora.

Lexical bundles have been retrieved using concordancing software called N-gram Statistical Package (NSP). This program, in fact, can find and display all bundles of different length in corpora with their actual frequencies.

Regarding syntactical structure, a modified version of Biber et al.'s structural framework (1999) is used for structural grouping of bundles found in the corpus. In other words, Biber et al.'s (1999) two broad groupings are maintained, but at the same time a four-division classification(Noun phrasal, prepositional phrasal, verb phrasal, and clausal bundles) has been employed for the present study in order to more accurately reflect the structural patterns of the lexical bundles in the *research corpus*.

5. Results and Discussion

After the application of the identification criteria, a total of 257 different lexical bundles, a list comprising 155 three-word and 102 four-word bundles, were identified in *research corpus*. These 257 bundles amount to a total of 5410 tokens, which make up about four percent of the total words in our corpus. As can be expected, the list is largely composed of three-word bundles which account for about sixty percent of the total bundles in the corpus.

Table 5.1 shows the twenty-five most commonly used bundles in *research corpus* in order of frequency (the number of their occurrences in *research corpus)*. These top twenty-five bundles, in fact, amount to 2210 tokens, which constitute about forty percent of the total bundles in the corpus.

Table 5.1. *Top 25 Bundles in Research Corpus in Order of Frequency*

Rank	three-word bundles	Frequency	four-word bundles	Frequency
1	one of the	187	is one of the	59
2	in order to	102	on the basis of	50
3	the results of	76	one of the most	40
4	as well as	69	on the other hand	39
5	is one of	69	in Tehran stock exchange	38
6	the present study	66	the results show that	36
7	the effect of	65	of this article is	31
8	Tehran stock exchange	64	of the most important	30
9	the relationship between	60	as one of the	28
10	based on the	59	the aim of this	28
11	in this paper	54	one of the important	26
12	of this research	53	this article is to	24
13	results show that	52	the results of this	23
14	the basis of	51	of this research is	22
15	the present paper	51	the purpose of this	21
16	this article is	51	the science of history	21
17	in this study	50	of this study is	20
18	of this study	50	one of the main	18
19	on the basis	50	the results of the	18
20	on the other	50	in the field of	17
21	in this article	45	that there is a	17
22	the impact of	45	at the same time	16
23	of this article	43	is an attempt to	16
24	the other hand	40	of this paper is	15
25	the results show	40	results show that the	15

As shown in table 5.1, sixty percent of these bundles end in a function word, such as an article or a preposition (e.g., *one of the, the results of*). As Biber et al. (1999, p. 992) point out; this is a hallmark of the bundles used in academic writing.

Another important observation made through looking briefly at the table is that there are certain semantic and structural relationships between the keywords of the lexical bundles. The term *keyword* refers here to the word that carries the meaning of the entire bundles. For instance, there are the two bundles *the result of this* and *the results of this* that share the same noun keyword *result* but have singular and plural forms. Although we do not intend to consider semantic relations between the keywords of the bundles, we might find it worthwhile to mention one specific instance of a synonymy relation between the keywords of the two bundles *in this paper* and *in this article*.

5.1. Comparing Forms

Of the 257 lexical bundles, only ninety-two were identified in Hyland's (2004) research article corpus. These ninety-two bundles amount to 2370 tokens, which represent forty-three percent of all bundle tokens in *research corpus*, whereas in Hyland's (2004) corpus they account for 11266 tokens, which represent about fifty-three percent of all bundle tokens. This finding suggests clearly that these bundles occur less frequently in research corpus as compared as compared to Hyland's (2004) corpus.

A closer look at the frequencies of these ninety-two bundles in Hyland's (2004) corpus reveals that fifty-three out of them occur at least seventy-five times (i.e., over sixty times per million words as a measure of the most frequent bundles) which, in turn, corresponds to six instances or higher per 100,000 words. This is, therefore, the normalized frequency adopted to compare the frequencies of the common ninety-two bundles in both corpora. Applying the normalized frequency, these ninety-two bundles can be classified into four different categories as follows:

a) Bundles Occurring at Least Six Times Per 100,000 Words in the Two Corpora

Table 5.2 shows the bundles occurring at least six times per 100,000 words in both corpora in alphabetical order. The bundles that are underlined are those that are less frequent in research corpus, while bundles that are in bold are those that their relative frequencies are roughly similar. Also, the bundles neither underlined nor in bold are those that their relative frequencies in the two corpuses are significantly different from each other. As shown in the table, forty-five out of fifty one bundles are three-word sequences and the remaining ones are four-word bundles. With the exception of the four bundles shown in bold, the rest of the bundles represent differences between the two corpora in terms of frequency patterns. It can be observed that many of these bundles are used in greater amounts in *research corpus* suggesting an overuse of some certain bundles compared to Hyland's (2004) corpus. This tendency among Iranian scholars to overuse some specific types leads to unnecessary repetitiveness and deprives their writings of the phraseological richness characteristic of well-written academic prose.

Table 5.2 *Bundles Occurring at Least 6 Times Per 100,000 Words in the Two Corpora*

Rank	Lexical Bundle	Research Corpus	Hyland Corpus
		Relative Frequency	Relative Frequency
1	according to the	19.24	15.28
2	a number of	12.11	18.72
3	as a result	14.25	12.88
4	as a result of	7.81	6.00
5	as well as	49.18	28.32
6	**at the same**	12.11	11.92
7	at the same time	11.36	7.28
8	**a variety of**	9.97	9.28
9	based on the	42.05	21.76
10	can be used	7.83	9.04
11	**due to the**	24.23	24.08
12	in addition of	12.78	7.36
13	in order to	72.70	44.64
14	in terms of	23.52	32.72
15	**in terms of the**	6.39	6.00
16	in this study	35.63	19.36
17	is based on	17.10	6.00
18	most of the	13.49	17.01
19	one of the	133.28	29.68
20	on the basis	35.63	9.52
21	on the basis of	35.63	9.84
22	on the other	35.63	22.00
23	on the other hand	27.79	13.12
24	part of the	12.11	13.92
25	some of the	22.80	12.88
26	that there are	9.97	17.04
27	that there is	20.66	11.76
28	the basis of	36.35	10.96
29	the case of	9.97	17.04
30	the concept of	9.97	7.12
31	the effect of	46.32	11.12
32	the effects of	27.08	10.08
33	the end of	19.95	13.65
34	the fact that	13.49	17.12
35	the formation of	12.78	7.12
36	the form of	7.12	7.92
37	The importance of	16.39	9.52
38	the level of	10.68	7.84
39	the nature of	21.38	8.56
40	the number of	12.78	26.96
41	The other hand	28.51	19.84
42	the present study	47.04	14.88
43	the process of	19.95	6.96
44	there is a	26.37	16.18
45	there is no	17.81	18.72
46	the relationship between	42.76	13.25
47	the results of the	12.82	7.20
48	the role of	27.08	12.96
49	the same time	11.39	10.56
50	the use of	27.08	29.36
51	with respect to	17.10	10.56

b) Bundles occurring at least six times per 100,000 words only in Hyland corpus

There are only two bundles *the absence of* and *in other words* occurring at least six times per 100,000 words in Hyland corpus but are rarely used in *research corpus*, with the relative frequency of 5.69.

c) Bundles occurring at least six times per 100,000 words only in research corpus

The complete list of bundles occurring at least six times per 100,000 words only in *research corpus* is presented in table 5.3 in alphabetical order. As shown in the table, nineteen out of twenty-nine are four-word bundles and the remainingare three-word ones. A close look at the relative frequencies of the bundles shown in the table reveals that twenty one of them are used less frequently in Hyland's (2004) corpus in comparison to *research corpus*, appearing less than four times.

Table 5.3. *Bundles Occurring at Least 6 Times Per 100,000 Words only in Research Corpus*

Rank	Lexical Bundle	Research Corpus	Hyland Corpus
		Relative Frequency	Relative Frequency
1	as well as the	9.94	4.80
2	at the end of	9.94	5.84
3	at the time of	7.10	2.24
4	different types of	7.12	2.80
5	for the development of	6.39	2.88
6	important role in	7.12	2.64
7	in addition to the	6.39	2.96
8	in this paper	38.48	2.00
9	is one of the	42.05	5.68
10	of the present study	8.52	3.36
11	one of the most	28.51	2.88
12	on the one hand	7.81	3.12
13	results indicate that	14.96	2.48
14	results show that the	10.65	2.48
15	that there is a	12.07	3.12
16	the analysis of	16.39	1.92
17	the analysis of the	9.94	2.16
18	the basis of the	7.10	2.88
19	the end of the	8.52	4.72
20	the lack of	7.12	2.72
21	the purpose of the	9.94	3.76
22	there is no significant	8.52	2.08
23	the result of	20.66	4.80
24	the results of this	16.39	2.56
25	the structure of	16.39	3.60
26	the study of	19.24	5.92
27	the study of the	7.81	2.16
28	there was no	7.12	4.16
29	with respect to the	6.39	4.96

d) Bundles occurring less than six times per 100,000 words in both corpora

There are ten instances occurring less than six times per 100,000 words in the two corpora eight of which are four-word bundles and the remaining are three-word ones. Table 5.4 presents the complete list of these bundles. The five bundles in bold are those that occur with similar frequencies in both corpora.

Table 5.4. ***Bundles Occurring Less Than 6 Times Per 100,000 Words in Both Corpora***

Rank	Lexical Bundle	Research Corpus Relative Frequency	Hyland Corpus Relative Frequency
1	an important role	4.98	2.40
2	**an important role in**	4.97	5.60
3	in the absence of	4.26	3.12
4	in the form of	5.68	2.96
5	**in the process of**	4.97	4.48
6	is based on the	5.68	3.28
7	**the majority of**	4.98	5.20
8	**the performance of the**	4.97	5.44
9	**the relationship between the**	5.68	4.08
10	the time of the	4.26	2.16

As shown in tables 5.2 through 5.4, there are only nine bundles occurring with similar relative frequencies in the two corpora. Furthermore, it can be observed that twenty-one out of twenty-nine most frequently bundles in *research corpus* appear less than four times per 100,000 words in Hyland corpus. Despite the fact that there are ninety-two common bundles found in both corpora, these findings, in general, show relatively large differences between the two corpora in the forms and frequencies of bundles employed.

5.2. Comparing Structures

Table 5.5 presents the structural classification of bundles identified in *research corpus* and their corresponding type and token frequencies.

Table 5.5 ***Frequency of Structural Categories of Bundles in Research Corpus***

Rank	Structural Patterns	Examples	Types	%	Tokens	%
A	**Noun Phrasal Bundles**					
1	Noun phrase with *of*-phrase fragment	one of the	85	33	1874	35
2	Complete Noun Phrase	the Persian gulf	12	5	332	6
3	Noun phrase with other post-modifier fragment	changes in the	10	4	160	3
B	**Prepositional Phrasal Bundles**					
4	Prepositional phrase with embedded *of*-phrase fragment	in terms of the	23	9	334	6
5	Other prepositional phrase (fragment)	on the other hand	55	22	1261	23
C	**Verb Phrasal Bundles**					
6	Passive verbs + prepositional phrase fragment	is based on the	4	1	102	2
7	Other passive structures	can be used	5	2	55	1
8	Copula *be* + [noun phrase/adjective phrase]	is one of the	7	3	201	3
9	[Dummy there/pronoun/noun phrase] + be + noun phrase fragment	there is a	11	4	266	5
10	Anticipatory *it* + verb phrase	it is argued	3	1	46	1
11	Noun phrase/noun + (verb) + that	results show that	12	5	225	4
12	Noun phrase + be + to	this article to	5	2	64	1
13	Other verb structures	deals with the	15	6	211	4
D	**Clausal Bundles**					
14	*That*-clause fragment	That there is a	4	1	69	1
15	Adverbial-clause fragment	as compared to	2	1	15	1
16	**Other Expressions**	as well as	4	1	195	4
Total			257	100	5410	100

When Biber et al.'s structural framework (1999) was applied to the bundles in *research corpus* (see table 3.1), it was found that their structural patterns covered most of these bundles' structural correlates. Only five new patterns were added to the original classification scheme underlined in the table.

It can be seen from table 5.5 above that the noun phrase with *of*-phrase fragment is the most common structure in *research corpus*. Together with noun phrases with other post-modifier fragments and complete noun phrases, they comprise over forty percent of all types and tokens in the corpus. This result coincides with recent findings and supports the view of academic writing as being 'noun-centric' (Swales 2008:5).

It can also be observed that adverbial-clause and *that*-clause fragments together with other expressions are the least frequently structures in *research corpus*.

Applying a modified version of a modified version of Biber et al.'s structural framework (1999), these sixteen structural patterns can be classified into four boarder categories as follows; "noun phrasal bundles" (NP-based), "prepositional phrasal bundles" (PP-based), "verb phrasal bundles" (VP-based), and "clausal bundles" (Cf-based). Of course, there are some bundles like *as well as the* thatdo not fit into four previously described categories. Figure 5.1 and 5.2 show the distribution of four different categories of lexical bundles in *research corpus*.

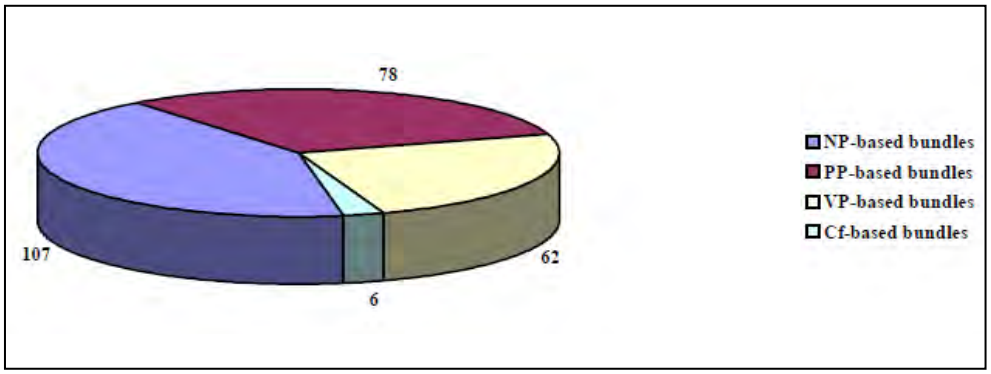

Figure 5.1. *Distribution of Four Categories of Lexical Bundles in Research Corpus by Type*

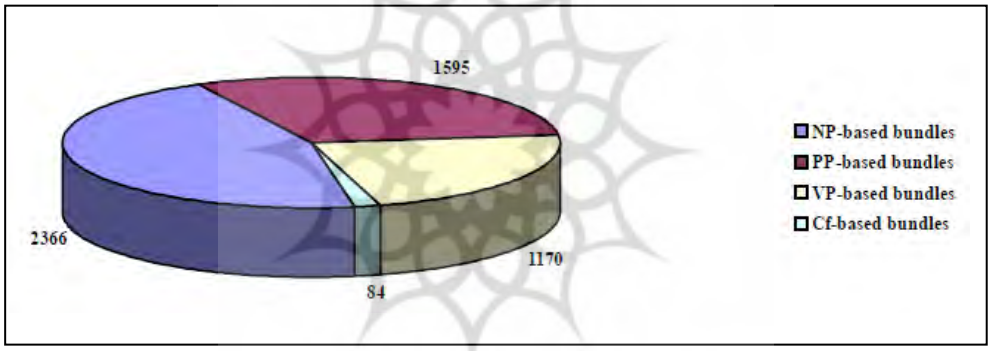

Figure 5.2. *Distribution of Four Categories of Lexical Bundles in Research Corpus by Token*

As illustrated by the figures above, over the third fourths of the bundles in the corpus begin with a noun or prepositional phrase (e.g., *the results of the*, *on the basis of*). This is in line with Biber et al.'s (2004) findings that almost seventy percent of the bundles in academic writing consist of noun phrase expressions or a sequence that bridges across two prepositional phrases (Biber et al., 2004, p.382).

Now, it is reasonable to compare the structures of the lexical bundles in *research corpus* to those found in Hyland's (2004) research article corpus. Table 5.6 presents the percentages of main structures of bundles in Hyland corpus.

Table 5.6. *The Percentages of Main Structures of Bundles in Hyland Corpus* (Hyland, 2008, p.10)

Rank	Structures	%
1	Noun phrase + *of*-phrase fragment	24.4
2	Passive verbs + prepositional phrase fragment	19.3
3	Other prepositional phrase (fragment)	17.5
4	Prepositional phrase + *of*-phrase fragment	13.5
5	Noun phrase with other post-modifier fragment	10.6
6	Anticipatory *it* + verb phrase	2.5
7	Others	9.5

As shown in tables 5.5 and 5.6, the structural comparison of bundles in *research corpus* and those found in Hyland's (2004) research article corpus reveals the following patterns.

First, just like Hyland corpus, *research corpus* is dominated by the bundles that incorporate a noun phrase in their structures which account for forty-two percent of all bundle types, and particularly by the noun phrase with *of*-phrase fragment. This structure comprises about thirty-three percent of all bundle types, a larger percentage than that represented by the same structure in native academic writing.

Second, the bundles beginning with a prepositional phrase in their structures are the second-ranked structural pattern, which account for thirty-one percent of all bundle types in both corpora.

Third, at first glance, it seems that there are remarkable similarities between the two corpora in the frequencies of noun phrasal and prepositional phrasal bundles employed. However, a closer look at the distribution of these structural categories in the two corpora reveals the apparent overuse of lexical bundles representing complete structural units in *research corpus*. This structure comprises about thirteen percent of all bundle types employed in *research corpus*, while, as mentioned previously, lexical bundles represent complete structural units occurring

rarely in the academic section of Biber et al.'s corpus (1999). For instance, the bundles *the present study* and *on the other hand*representing complete structural units occur 47.04 and 27.79 times per 100,000 words in *research corpus* respectively, while they appear 14.88 and 13.12 times in Hyland corpus.

Fourth, one remarkable difference between the two corpora is the significant underuse of lexical bundles comprising passive verbs in *research corpus*. As shown in table 5.5, passive bundles comprise 19.3% of all bundles employed Hyland corpus but only three percent of all bundle types and tokens in *research corpus*. As Hyland (2008, p.11) points out, passive bundles are employed in the discussion of research methods and logical reasoning, so as to depersonalize these statements and make them more objective and universal. On the other hand, the bundles with anticipatory *it* usually are used to communicate the writer' stance by presenting the proposition as an obvious and widely accepted fact. In the case of bundles beginning with anticipatory *it*, as shown in table 5.5, this structure is rarely employed in *research corpus*. These findings, altogether, demonstrate Iranian writers' inability to employ passive structures and impersonal form in the construction of convincing argument.

6. Concluding Remarks

Since this study is concerned with non-native professional writers who have as much experience and knowledge of their discipline as the native scholars to whom they are being compared, it is reasonably likely to find some similarities between the two corpora as far as the use of lexical bundles is concerned. However, the analysis of the use of lexical bundles reveals that there are relatively large differences between the two corpora in the forms and frequencies of bundles employed.

There are two important differences that deserve to be underscored here. First is the abundant use of bundles in *research corpus* that could be taken as the most

surprising result given the fact that the previous research shows that less proficient language users, especially non native writers tend to rely less on bundles in the development of their discourses. However, further examination of the use of bundles in *research corpus* show the lesser degree of variety in Iranian scholars' writing brought about bytheir excessive reliance on a handful of highly frequent bundles which feel confident in using. This pattern of overuse can contribute to a certain degree of repetitiveness and lack of stylistic variety in non-native writing.

Second, the limited use of passive bundles in *research corpus* shows Iranian writers' difficulties with this particular structure. This is very important considering that this structure and also the bundles with anticipatory *it* constitute a more subtle aspect of academicwriting. In fact, the expression of writer stance, the delicate engagement and persuasion of the reader and the proper manipulation of impersonal forms are strategies that academic writers must master if they are to be successful in disseminating their work to the larger scientific circles. Passive bundles, among other things, play an essential role in this process. This finding emphasize the need for the explicit teaching of passive bundles in academic writing, as their use proves to be a complicated task even for non-native professional authors.

Of course, there are some reasons for caution. To start with, some differences between the two corpora may be likely related to the limited size of our corpus. Clearly, additional benefits could be obtained from the analysis of a larger corpus. Furthermore, the analysis of motivations behind the selection of lexical bundles is still required.

Although more research is still needed, the findings can be an important contribution to pedagogic practice and offer useful insights for teachers and materials writers. In short, learning the appropriate ways of using the most preferred word combinations in academic writing can certainly help Iranian writers to produce a natural text and to write more effectively.

References

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

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

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, *23*(4), 397-423.

Cortes, V. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3, 43-57.

Firth, J. R. (1964). Papers in Linguistics. London and New York: OxfordUniversity Press.

Haswell, R. (1991). Gaining ground in college writing: Tales of development and interpretation. Dallas: Southern MethodistUniversity Press.

Hyland, K. (2004). Disciplinary discourse: Social interactions in academic writing, Ann Arbor: University of Michigan Press.

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

Jalali, H. (2009). Lexical bundles in Applied Linguistics: Variations within a single discipline (Unpublished doctoral dissertation). Isfahan University, Isfahan.

Modarres Khiyabani, SH. (2007). Investigating lexical collocation in Persian (Unpublished doctoral dissertation). Allame Tabatabayee University, Tehran.

Radford, A. (2004). English syntax: An introduction. Cambridge University Press.

Simpson-Vlach, R., & Ellis, N.C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487-512.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Valipoor, L. (2010). A corpus-based study of words and bundles in chemistry research articles (Unpublished Master's thesis). Kashan University, Kashan.