International Journal of Information Science and Management
Vol. 13, No. 1, 2015, 51-69

Structural and Functional Analysis of Lexical Bundles in Medical Research Articles: A Corpus-Based Study

Zahra Sadat Jalali
English Department
Kashan University
Z.S.Jalali25@gmail.com

Mohammad Raouf Moini
English Department
Kashan University
rmoin@kashanu.ac.ir

Mohamad Alaee Arani
Knowledge and information Science
Payame Noor University ,Mashhad, Iran
Corresponding author: alaee62@gmail.com

Abstract

As a member of larger family of formulaic sequences**,** lexical bundles play different discourse functions in written research articles. This study investigated the use of four-word lexical bundles in published research articles in medicine via natural language processing by computational linguistics. A corpus of 2,420,914 words was extracted from 790 research articles in 33 medical disciplines. For the identification of lexical bundles, a number of computer software products such as ABBYY FineReader 10 professional edition, Total assistant, Antconc 3.2.3, and WordSmith Tools 5 were used. The identified lexical bundles were classified structurally and functionally based on the taxonomies in the literature. The results of the study showed that 102 identified lexical bundles differ structurally and functionally and most of the writers of medical research articles rely on text-oriented bundles for establishing their written academic discourse. This study provided new insights in understanding the discipline-specific discourse of medical research articles and in doing further corpus-based research in written academic discourse and EAP. This research introduced stylistic linguistics point of view in information retrieval systems development.

Keywords: Lexical bundles, Research article, Corpora, Medicine, Natural language processing, Computational linguistics

Introduction

Computational linguistics is an interdisciplinary field addressing human languages by applying methods of not only linguistics but also computer and information sciences (Hyland, 2008b). Research in computational linguistics addresses the computational properties of linguistic models of natural language and develops algorithms and computational implementations of such linguistic models. Research in Natural Language Processing (NLP) also emphasizes the goal of upward systems that can deal effectively with natural language data in academic contexts such as English for Academic Purposes (EAP). These types of

research focus on a particular NLP application type, on language technological attitude, on algorithmic techniques, and on a linguistic formalism applied in computational linguistics.

In the simplest form of automatic text retrieval, users enter a string of keywords that are used to search the inverted indexes of the document keywords. This approach retrieves documents based solely on the presence or absence of exact single word strings as specified by the logical representation of the query. This approach will miss many relevant documents because it does not capture the complete or deep meaning of the user's query. However, recent corpus-based studies have found that there are EAP-specific word combinations that are semantically and syntactically compositional (Biber, Johansson, Leech, Conrad, & Finegan, 1999; Biber, 2004). In fact, these word combinations are built based on specific EAP or some technical words which fulfill rhetorical and discourse-related functions implemented in retrieval systems and prominent in academic writing such as introducing and elaborating topics, hypothesizing, concluding, summarizing, etc. (Karlgren, 2000).

The history of formulaic patterns in applied linguistics dates back to Jespersen (1924) and Firth (1951), who popularized the term "collocation". Since late 1970, much more attention has been paid to formulaic sequences for language processing and production (Hakuta, 1974; Nattinger & DeCarrico, 1992; Wray, 2002). Being defined as multi-word combinations that are stored and retrieved holistically from the mental lexicon upon speech, formulaic sequences have been considered to minimize encoding work for the speaker and decoding work for the addressee, thus allowing for the construction of fluent spoken discourse (Erman, 2007; Wood, 2006). Proper use of formulaic sequences has also been found to be critical for the acquisition of native-like language competence (Dufon, 1995; House, 1996). Reviewing the related literature, it was found that there are different terms that are used to refer to multi-word combinations. These terms are *clusters* (Hyland, 2008a; Schmitt, Grandage & Adolphs, 2004), *recurrent word combinations* (Altenberg, 1998; De Cock, 1998), *phrasicon* (De Cock, Granger, Leech, & McEnery, 1998), *n-grams* (Stubbs, 2007a, 2007b), and *lexical bundles* (Biber & Barbieri, 2007; Cortes, 2002).

In recent years, corpus linguistic studies or corpus-based research has shed light on distinctive linguistic features of academic discourses. Corpus linguistics focuses the term "lexical bundle". According to Biber, Johansson, Leech, Conrad and Finegan (1999), it first appeared in the Longman Grammar of Spoken and Written English. However, the concept of lexical bundle dates back to Salem (1987). He carried out a research on the analysis of a corpus of French government documents and texts.

As Biber and Barbieri (2007) mentioned, lexical bundles are not structurally complete and they are not idiomatic in meaning but they serve important discourse functions in both spoken and written texts. Cortes (2002) presents another feature for lexical bundles and believed that idiomaticity and fixedness are qualities used to describe lexical bundles. Hyland (2008a) mentioned that lexical bundles are register-specific and change from one discipline or register

to another.

Conducting a series of studies in the field of lexical bundles and comparing bundles across registers, Biber et al. (1999) found that grammatical structure of lexical bundles is a distinct characteristic of registers. Some studies have investigated the similarities and differences of lexical bundles across different genres within one discipline (Cortes, 2004; Hyland, 2008a; Jalali, 2009; Valipoor, 2010; Parvizi, 2011). 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. Some of the studies worked on the structure of lexical bundles in text sections and compared rhetorical functions that they serve in those sections (Martinez, 2003; Valipoor, 2010; Parvizi, 2011). For instance, Valipoor (2010) identified lexical bundles in the genre 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.

Reviewing studies done on medical research articles revealed that there are studies which have presented two-word collocations (Marco, 2000) and word lists (Wang, Liang & Ge, 2008), but no study has been done on lexical bundles in medical research articles. For instance, Marco (2000) worked on the linguistic pattern and collocation frameworks selected by a specific genre. He found that the collocation selection for these frameworks was based on the linguistic conventions of the genre.

The main objective of the present study was to identify the four-word lexical bundles in published medical articles. As different academic discourses rely on different repertoires of lexical bundles, readers or writers of such articles should be aware of these lexical bundles in order to be more competent in their recognition, comprehension, and production. This knowledge is achieved through knowing a list of these bundles, being exposed to texts which include these bundles and practicing and using them. Furthermore, forms of lexical bundles used in research articles and functions they play were investigated.

In order to reach a comprehensive analysis of lexical bundles used in published medical research articles, this study will explore the following research questions:

1. What are the most frequent lexical bundles in medical research articles (MRAs)?
2. What are the forms of lexical bundles used in MRAs?
3. What functions do lexical bundles play in MRAs?

Methodology

To collect the required RAs for establishing the corpus (Corpus of Medical Research Articles, referred to as COMRA hereafter in this study), the Science Direct Online (SDO)[1] was used. The database SDO is considered to be one of the most authoritative and representative databases. All the research articles in medicine which were adopted in this corpus were downloaded from an authentic database, SDO.

In the discipline of Medicine and Dentistry, there are 33 subject areas. Following Wang, Liang and Ge (2008), in this study all areas of medical sciences were included. All journals in the 33 subject areas published during 2009-2011 were used for compilation of the corpus. In each year, two issues of each volume were randomly selected (cluster-random sampling) and only downloaded articles that followed the (Introduction, Method, Result and Discussion) structure were included in the study. Table 1 shows the selected medical subject areas. About 21 articles[2] in each of the 33 subject areas were selected while each article on average included about 3000 words. Ultimately, 790 articles were compiled in order to produce the corpus of 2,420,914 words.

Table 1

Medical subject areas

Anesthesiology and Pain Medicine (General)	Medicine and Dentistry
Cardiology and Cardiovascular Medicine	Nephrology
Clinical Neurology	Obstetrics, Gynecology and Women's Health
Complementary and Alternative Medicine	Oncology
Critical Care and Intensive Care Medicine	Ophthalmology
Dentistry, Oral Surgery and Medicine	Orthopedics, Sports Medicine and Rehabilitation
Dermatology	Otorhinolaryngology and Facial Plastic Surgery
Emergency Medicine	Pathology and Medical Technology
Endocrinology, Diabetes and Metabolism	Perinatology, Pediatrics and Child Health
Forensic Medicine	Psychiatry and Mental Health
Gastroenterology	Public Health and Health Policy
Geriatrics and Gerontology	Pulmonary and Respiratory Medicine
Health Informatics	Radiology and Imaging
Hematology	Surgery
Hepatology	Transplantation
Immunology,Allergology& Rheumatology	Urology
Infectious Diseases	

All the medical research articles included in this corpus were kept in their original length and were written in the internationally conventionalized IMRAD structure.

In terms of the size of the corpus, we followed the principle suggested by Biber (2006). He argues that a corpus must be large enough to decently represent the occurrence of the

features being studied. He also explains the importance of corpus size emphasizing that it depends on the purpose of the study. Consequently, for the present study a written specialized corpus containing 2,420,914 running words from 790 written texts of a single genre (medical research articles) was produced so that each subject area encompassed an equal number of 21 articles while each article included 6 pages on average and about 3000 words. After collecting electronic files, the process of standardization which is erasing non-textual annotations such as titles, the charts, diagrams, bibliographies, tables, page numbers, formulations and pictures was completed in order to produce files being readable by computer programs utilized in this research.

A frequency of 20 times per million words corpus with a requirement that this rate of occurrence be realized in at least five different texts were considered as criteria. Identification of 4-word lexical bundles is the center of attention in the COMRA because 4-word bundles are far more common than 5-word strings and offer a clearer range of structures and functions than 3-word bundles (Hyland, 2008a). Furthermore, many 4-word strings hold 3-word bundles in their structure (Cortes, 2004). After data had been collected, three computer software programs were applied in order to identify lexical bundles. The following section introduces these computer programs.

Computer and software programs

The computer software used in this study included ABBYY FineReader 10 professional edition, Antconc 3.2.3, and WordSmith Tools 5. ABBYY FineReader is an Optical Character Recognition (OCR) system and intelligent document processing software which is used to convert scanned documents, PDF files and documents and image files into editable format.

Application of ABBYY FineReader allows producing plain texts which can be uploaded to Antconc (Anthony, 2007). The concordance tool of Antconc software was used and files were given to this software and the cluster size of 4-word (for min and max size) was counted. Then, different keywords or search terms such as articles, to be verbs, modals, prepositions, and demonstrative adjectives were typed. Also, a cut-off frequency of 20 per one million words was set. As a consequence, the minimum cluster frequency of 48 for a corpus of 2,420,914 was given to Antconc software. Then, this software displayed clusters of words that surrounded a search term and ordered them alphabetically or by frequency. Like Antconc, WordSmith (Scott, 2008) was used to extract and identify lexical bundles in different texts. The WordSmith has the additional advantage of showing the number of texts in which lexical bundles happen.

The next stage was the structural and functional classification of the lexical bundles. The former was based on Biber et al.'s (1999) structural taxonomy and for the latter Hyland's (2008a) functional taxonomy was used. Hyland's taxonomy is based on academic registers. Since the focus of this study was on a specific academic register of research articles, this kind

of taxonomy was used in functional classification of bundles. The general categories in this taxonomy are:

Research-oriented: Help writers to structure their activities and experiences of the real world. The sub-categories of lexical bundles in this group are as follows:

- **Location**- indicating time and place, e.g. *in the present study.*
- **Procedure**- indicating methodology or purpose of research, e.g. *the purpose of this.*
- **Quantification**- describing the amount or number, e.g. *is one of the.*
- **Description**- detailing qualities or properties of material, e.g. *in the control group.*
- **Topic**- related to the field of research, e.g. *in the United States.*

Text-oriented: These clusters are concerned with the organization of the text and the meaning of its elements as a message or argument and include:

- **Transition signals**- establishing additive or contrastive links between elements, e.g. *on the other hand, as well as the.*
- **Resultative signals**- mark inferential or causative relations between elements, e.g. *the results of the.*
- **Structuring signals**- text-reflexive markers which organize stretches of discourse or direct readers elsewhere in the text, e.g. *as shown in fig*.
- **Framing signals**- situate arguments by specifying limiting conditions, e.g. *in the presence of.*

Participant-oriented: These are focused on the writer or reader of the text. Sub-categories of participant-oriented bundles are:

- **Stance features**- convey the writers' attitudes and evaluations. According to Cortes (2002), this category includes attitude markers, epistemic-certain, epistemic-uncertain and intention bundles, e.g. *were more likely to*.
- **Engagement features**- address readers directly, e.g. *it should be noted*.

Results

Application of the criteria proposed by Biber et al. (1999) about the identification of lexical bundles and utilization of disparate computer programs yielded 102 different 4-word lexical bundles in the full corpus of 2,420,914 words of published medical research articles. Table 2 represents all these lexical bundles in the COMRA.

Table 2

Lexical bundles in COMRA

	Lexical bundles	Frequency	No. of texts		Lexical bundles	Frequency	No. of texts
1	in the present study	453	220	52	may be due to	72	64
2	on the other hand	258	182	53	with the exception of	72	57
3	in the presence of	224	120	54	are more likely to	72	49
4	At the end of	201	119	55	for each of the	70	57
5	at the time of	185	123	56	as a function of	69	31
6	were more likely to	177	77	57	an increase in the	68	51
7	on the basis of	166	100	58	results of this study	67	55
8	the end of the	165	112	59	for the treatment of	67	54
9	It is possible that	163	113	60	in the treatment of	66	49
10	as well as the	159	125	61	presence or absence of	65	49
11	The results of the	148	117	62	are presented in Table	64	53
12	of the present study	148	112	63	was used to determine	64	57
13	as shown in Fig	148	80	64	aim of this study	63	62
14	in the control group	139	65	65	are summarized in Table	63	52
15	In the current study	127	67	66	were obtained from the	63	57
16	are shown in Table	126	101	67	Are summarized in table	63	52
17	this study was to	125	116	68	and the number of	61	48
18	It is important to	123	100	69	in terms of the	61	41
28	studies have shown that	104	89	79	In this study were	61	57
19	in the absence of	121	84	70	as well as in	60	56
20	In the case of	121	82	71	has been reported to	60	49
21	more likely to be	119	61	72	In this study, the	59	52
22	been shown to be	118	96	73	the fact that the	59	51
23	was found to be	117	88	74	with the use of	59	46
24	in the United States	109	77	75	the extent to which	58	35
25	is one of the	108	91	76	as compared to the	58	25
26	In this study, we	107	93	77	purpose of this study	56	53
27	an important role in	106	86	78	In our study, the	56	52
29	as a result of	103	84	80	present study was to	55	54
30	have been shown to	98	75	81	important role in the	55	50
31	one of the most	94	86	82	may be related to	55	45
32	As shown in Table	93	66	83	It should be noted	٥٥	43
33	The results of this	92	75	84	the basis of the	55	42
34	in accordance with the	85	75	85	the presence or absence	54	40
35	The purpose of this	82	74	86	a large number of	53	46

	Lexical bundles	Frequency	No. of texts		Lexical bundles	Frequency	No. of texts
36	at the same time	82	70	87	by the fact that	52	47
37	the total number of	81	62	88	is consistent with the	52	45
38	It has been shown	81	67	89	were less likely to	52	31
39	in the context of	79	55	90	the presence of the	51	38
40	was used as a	77	66	91	a role in the	51	45
41	were found to be	77	57	92	at the beginning of	50	41
42	was defined as the	77	63	93	higher than that of	50	31
43	a wide range of	76	66	94	could be due to	50	27
44	with respect to the	76	62	95	play an important role	49	45
45	during the study period	76	43	96	was obtained from the	49	43
46	The number of patients	75	43	97	the size of the	49	40
47	in the number of	75	57	98	The time of the	48	42
48	In addition to the	74	69	99	plays an important role	48	45
49	Be due to the	74	64	100	should be noted that	48	40
50	test was used to	73	66	101	the presence of a	48	39
51	The aim of this	72	71	102	was used for the	48	47

As can be seen, there were 102 different lexical bundles in the COMRA which was a relatively large corpus of more than two million words. The results of the current study showed that just 0.3% of the whole corpus consisted of lexical bundles. As shown in table 2, the most frequent lexical bundle is *in the present study* with the frequency of 453 in the corpus occurring in 220 texts. This high frequency indicates that in each one million word corpus this bundle has occurred about 226 times, which is 11 times more than the frequency of 20 per each one million word. Unlike the highest frequent lexical bundles, *the time of the, plays an important role, should be noted that, the presence of a,* and *was used for the* are the least frequent lexical bundles in the COMRA. In addition, the number of texts in which these lexical bundles occurred in was very low (table 2). These bundles with the frequency of 48 occurred at least in 39 texts which show that these lexical bundles have occurred 9 times less than the most frequent lexical bundle.

Being identified on the basis of their frequency, lexical bundles were classified structurally. In this study, just 9 major structural categories were distinguished (table 3). The lexical bundles were classified based on whether they had started with nouns, prepositions or verbs.

Structural classification of bundles

Identified lexical bundles were classified into the taxonomy proposed by Biber, Johansson, Leech, Conrad and Finegan (1999). The categorization of lexical bundles revealed

that the largest structural category of lexical bundles was prepositional phrases, making up about 44.5% (with and without "of") of the total number of lexical bundles. Noun phrases with the overall frequency of 1842 (about 20.42%) formed another group of bundles of the whole corpus. The least frequent group of bundles was verb phrase+that clause fragments which formed about 1.7% of the bundles. The structural classification of lexical bundles is presented in table 3.

Table 3

Structural classification of lexical bundles in COMRA

Structures	examples	No. of bundles	Overall frequency	Percentage
Noun phrase+ of	*The end of the, the results of the*	19	1391	15.42
Other noun phrase	*A role in the, the extent to which*	7	451	5.00
Prepositional phrase+ of	*In the presence of, in the development of*	18	1850	20.50
Other prepositional phrases	*At the same time, between the two groups*	19	2168	24.03
Passive+ prepositional phrase fragment	*Are shown in table, was used for the*	13	1014	11.24
Anticipatory it+ verb/adjective	*It is possible that, it is important to*	4	422	4.7
Be+ noun/adjectival phrase	*Is one of the, is consistent with the*	3	234	2.59
Verb phrase+ that clause fragment	*Should be noted that, studies have shown that*	2	152	1.7
Verb/adjective+ to-clause fragment	*Are more likely to, can be used to*	3	301	3.33
Adverbial clause fragment	*As shown in figure, as compared to the*	3	299	3.31
Others	*This study was to, test was used to, as well as in*	11	738	8.18
Total		**102**	**9020**	**100**

As demonstrated in table 3, in the COMRA about 64.95% of lexical bundles are phrasal among which prepositional phrases form the most frequent clusters. As shown, about 26.87% are clausal bundles in research articles and among clausal bundles those beginning with passive+prepositional phrase fragments are more frequent than other groups.

As previously mentioned, the most frequently used lexical bundles were prepositional phrases in research articles to identify a particular time period or location. Clusters such as *in the present study, on the other hand, in the presence of, at the end of, and at the time of* had the highest frequency in the corpus.

In the classification of verb-phrases which included sub-categories such as passive+ prepositional phrase, anticipatory it+ verb/adjective, copula be+ noun/adjectival phrase, verb phrase+ that clause fragment, and verb/adjective+ to-clause fragment, the passive forms achieved the highest rank or overall frequency of 1014. In addition to phrasal and clausal fragments, there is another group of lexical bundles which is called by Biber et al. (1999) as "lexical bundles that do not fit neatly into any of other categories" (p.1024). These bundles formed just 8.18% of the whole bundles identified in this study, for example *as well as the, study was approved by, consent was obtained from.*

Functional classification of lexical bundles

In addition to structural or grammatical classification of lexical bundles, it is useful to classify them according to their function or meaning since lexical bundles tend to have functional characteristics that represent a register in which they are found (Biber, Conrad & Cortes, 2004; Biber et al., 1999; Hyland, 2008a).

As mentioned in methodology section, the taxonomy used for the functional analysis of lexical bundles was developed by Hyland (2008a) and included three major categorizations; research-oriented bundles, text-oriented bundles and participant-oriented bundles with various sub-categories for each. Table 4 presents the results of functional classification of the bundles. This includes the number of bundles, their frequency and the number of texts in which they are applied specifically to each category and sub-categories.

Table 4

Functional classification of lexical bundles in COMRA

Type of bundles	No. of bundles	Frequency	Percentage
Research-oriented bundles	37	**3347**	36.53
Location	12	1655	18.06
procedure	12	747	8.15
Quantification	10	648	7.07
Description	2	188	2.05
Topic	1	109	1.18
Text-oriented bundles	40	**3892**	42.47
Transition signals	4	551	6.01
Resultative signals	12	1122	12.24

Type of bundles	No. of bundles	Frequency	Percentage
Structuring signals	7	652	7.11
Framing signals	17	1567	17.10
Participant-oriented bundles	25	**1923**	21
Stance features	23	**1820**	19.86
Attitude markers	5	381	4.15
Epistemic-certain	3	163	1.77
Epistemic- uncertain	10	906	9.88
Intention	5	370	4.03
Engagement features	2	**103**	1.12
Total	**102**	**9162**	**100**

As shown in table 4., about 36.5% of the bundles belong to research-oriented bundles used to describe time, place, size and magnitude, the study itself, and research procedures in academic texts. As mentioned before, the most frequent lexical bundle is *in the present study* which forms 0.98% (about 0.1%) of the overall bundles. Although the most frequent bundles in the corpus is placed in the category of research-oriented bundles, it can be claimed that these medical research articles are characterized by a heavy use of text-oriented clusters especially framing signals and low use of participant-oriented bundles, because text-oriented bundles form about 42.5% of the whole bundles in COMRA in which framing signals with frequency of 1567 form the highest frequently used lexical bundles.

Participant-oriented bundles had the lowest frequency. These bundles focus on the writer or reader of the text. These bundles form about 21% of the COMRA among which epistemic-uncertain bundles were the most frequent clusters. Engagement lexical bundles are used to engage readers, e.g. *should be noted that*. As shown in table 3, these bundles made about 1% of MRAs. The purpose of these bundles is to direct the readers to certain understanding and lead them to particular interpretation.

Discussion

In recent years, more studies have been conducted in the area of corpus linguistics and formulaic sequences (Hyland, 2008a; 2008b, Jalali, 2009; Valipoor, 2010; Parvizi, 2011). Most of the studies have shown that these formulaic sequences- an umbrella term for lexical bundles- can be different from one discipline to another. According to these studies, it can be argued that these lexical bundles have different structures and functions based on the context in which they are used. In fact, the context establishes the function of bundles.

The results of this study showed that in COMRA, 102 four-word lexical bundles were the most frequent bundles while in studies carried out by Jalali (2009), Valipoor (2010) and

Parvizi (2011), the number of bundles is different. For example, Jalali identified 121, 255 and 141 bundles in the three corpora of research articles, master dissertations and doctoral theses in applied linguistics, respectively; while Valipoor found that there were just 223 bundles in the 4,000,000 word corpus of chemical research articles (CRAC). In her study, Parvizi (2011) found that there were just 24 bundles in a 2 million word corpus in the field of education. A comparison between the first twenty bundles in the present study and other three studies revealed that the common bundles between them were: *on the other hand, on the basis of, as well as the* with different frequencies and those which were available in the present study but not in other three studies were: *of the present study, in the control group, in the current study, are shown in table, this study was to*. Another comparison between the above mentioned number of bundles of the current study and the study carried out by Hyland (2008a) showed bundles which were not common: *at the time of, were more likely to, in control group, in the current study, are shown in table, his study was to, in the absence of.* Chen and Baker (2010) carried out a study on lexical bundles in L1 and L2 academic writing in which three corpora of FLOB-J, BAWE-EN and BAWE-CH representing native expert writing, native peer writing (produced by peer L1 English students) and learner writing (contained essays produced by L1 Chinese students of L2 English) were included. Again a comparison between the first twenty lexical bundles of the current study and this study was done and the same results were obtained. All in all, it can be concluded that some of the lexical bundles are discipline-specific because they are identified in this but not in other studies: *in the present study, at the time of, were more likely to, the results of the, of the present study, as shown in fig, in the control group, in the current study, are shown in table, this study was to, it is important to*.

Regarding the frequency of the bundles, in the present study, the frequency of about 28.5% of lexical bundles was more than 100. Identifying 223 bundles in her study, Valipoor (2010) showed that 71% of the bundles had a frequency of more than 100. The results of the study done in the area of applied linguistics by Jalali (2009) indicated that in genres of research articles, master theses and doctoral dissertation, 4%, 0.78%, and 2% of the bundles had a frequency of over 100, respectively. The results of the study done by Parvizi (2011) demonstrated that 67% of the bundles had a frequency of over 100.

In order to answer the second and third questions of this study, these lexical bundles were classified structurally and functionally. Findings of the study revealed that prepositional phrases were the most frequently used lexical bundles in medical research articles structurally. This result is exactly in line with the finding of the study carried out by Hyland (2008a), who found the overall frequency and percentage of phrasal lexical bundles more than clausal bundles, lending support to the idea or findings of previous studies such as Biber et al. (1999) who found that in academic writing most of the lexical bundles are phrasal rather than clausal. Comparison between studies on structures of lexical bundles is presented in table 5.

Table 5

Structural comparison of bundles in different disciplines

Research studies / Structures of LB	Hyland biology	Hyland Electrical engineering	Hyland Applied linguistics	Hyland Business studies	Jalali research articles	Jalali Master dissertation	Jalali Doctoral theses	Valipoor CRAC	Parvizi education	Chen and Baker FLOB-J	Chen and Baker BAWE-EN	Chen and Baker BAWE-CH	Jalali medicine
Noun phrase+ of	23.7	22.3	22.9	28.5	23.45	25.23	25.92	18.73	17.18	32.5	15.4	15	15.42
Other noun phrase	-	-	-	-	10.05	11.19	8.90	2.60	10.37	-	-	-	5.00
Prepositional phrase+ of	9.2	7.9	19.9	16.00	30.47	15.66	24.38	22.25	32.07	-	-	-	20.50
Other prepositional phrases	13.7	11.6	24.4	19.7	19.00	24	21.92	11.88	24.46	36	28.8	32.5	24.3
Passive+ prepositional phrase fragment	31.3	29.8	6.9	9.00	2.4	2.62	2.3	20.9	-	7	10.6	5	11.24
Anticipatory it+ verb/adjective	6.3	8.4	5.6	4.5	5.6	1.88	2.1	4.09	-	8.8	5.8	8.8	4.7
Be+ noun/adjectival phrase	-	-	-	-	1.22	1.72	1.8	3.15	-	2.6	10.6	6.3	2.59
Verb phrase+ that clause fragment	-	-	-	-	-	-	-	-	-	2.6	4.8	6.3	1.7
Verb/adjective+ to-clause fragment	-	-	-	-	-	-	-	-	-	7	18.3	15	3.33
Adverbial clause fragment	-	-	-	-	-	-	-	2.86	-	-	-	-	3.31
Others	6.4	9.2	10.7	9.9	7.81	17.7	12.68	7.07	15.9	2.6	2.8	4.8	8.18

Jalali (2009), Valipoor (2010) and Parvizi (2011) found that 75%, 55% and 84% of the bundles were phrasal, respectively. In these studies, like the present study, prepositional phrase+of was the most frequently used bundle in phrasal bundles as well. Besides, Chen and Baker (2010) showed that most of the native expert writers used a wide range of noun and prepositional phrases while English and Chinese students used more verb phrase bundles than expert writers did. Cortes (2004) revealed that bundles in history were mostly noun and prepositional phrases while in biology more structural categories were found. Also, a comparison between students and published writings indicated that most of the students rarely used bundles identified in published writings. The findings of his study were not statistically mentioned, so it was not possible to present them in the table.

Based on the results of the present study and regarding the functional analysis of lexical

bundles, the high application of text-oriented bundles can represent a sophisticated approach toward language. Most of specialists and scholars in medicine have used these kinds of bundles to show that they are competent academics. This is because they are experts in a specific medical area and they know their audiences. Furthermore, not only have they used these groups of lexical bundles to show the disciplinary competence but also to organize their discourse in the way that their readers have better understanding of the text.

Based on the findings of the study, about 17% of text-oriented bundles in medical sciences were to frame arguments, make connections, specify cases, and referred to limitations. It is worth mentioning that most of these framing signal bundles were made up of prepositional phrases+of. The results of this study are in agreement with findings of the study carried out by Hyland (2008b). He found that framing devices comprised a high proportion of text-oriented bundles. Based on his findings, writers in disciplines of applied linguistics and business studies mostly used text-oriented bundles. Table 6 presents the results of the functional comparison of bundles between disciplines briefly.

Table 6

Functional comparison of bundles between disciplines

	Hyland biology	Hyland Electrical engineering	Hyland Applied linguistics	Hyland Business studies	Jalali research articles	Jalali Master dissertation	Jalali Doctoral theses	Jalali medicine
Research-oriented	48.1	49.4	31.2	36	45.17	49.81	33.53	36.53
Text-oriented	43.5	40.4	49.5	48.4	41.95	35.73	53.83	42.47
Participant-oriented	8.4	9.2	18.6	16.6	12.88	14.46	12.64	21

On the contrary, the functional analysis of the study carried out by Jalali (2009) showed that 45% of the bundles were research-oriented bundles in the genre of research articles and master theses, while in the genre of doctoral dissertations he found that similar to our study, text-oriented bundles with a frequency of 1761 (54% of overall bundles) had priority over other functional categories. In another study, Chen and Baker (2010) found that both native and non-native students mostly used discourse/text organizers while native professional writers exhibited a wide range of referential markers. Nekrasova (2009) conducted a study on

the knowledge of English L1 and L2 speakers of lexical bundles and used structural and functional classification by Biber, Johansson, Leech, Conrad and Finegan (1999). He argued that in contrast with referential bundles, discourse-organizing bundles play a very important role in the comprehension of topic being discussed since they help speakers to develop the discourse and provide orientation for the listener as well. Moreover, Parvizi (2011) found that research-oriented bundles outweighed all other functional types of bundles while participant-oriented bundles were the least frequent functional type.

In the present study, the next most frequent group of lexical bundles was resultative signals which built about 12% of text-oriented bundles. Resultative markers reveal writers' interpretation of research processes and findings. These bundles play the role of rhetorical presentation of the research, since they present the conclusion to be drawn from the study and help writers show inferences which they want readers to draw from the discussion. The functional results of this study were in line with the study carried out by Jalali (2009) who discussed that in the category of text-oriented bundles, framing and resultative signals were the most frequent bundles in three genres of research articles, master theses, and doctoral dissertations. Regarding participant-oriented bundles, it is concluded that most MRA writers have used just two clusters (*it should be noted, should be noted that*) to engage readers in the text. They form about 21% of the whole bundles. It has been generally concurred that formulaic sequences such as *it should be noted that* and *as a result of* are central to the production of academic texts and discourses (Hyland, 2008b). This low percentage of the mentioned bundles shows that probably medical researchers are not aware of the function of these bundles or the influence that they have on the readers or they may use other bundles which are not highly frequent to convey their ideas.

The results of this study demonstrated that the frequency of the identified lexical bundles is really high and most of them happened in more than five different texts. As a consequence, it may be thought that it is not necessary to raise the awareness of EMP (English for Medical Purposes) students toward these clusters because they encounter them in the texts repeatedly, but some researchers have emphasized the fact that perceptual salience and developmental readiness are more important than frequency (Gass & Mackey, 2002). It means that in addition to frequency, learners should be aware of the function of lexical bundles. Schmidt (1990) argued that one useful way to help students get familiar with lexical bundles is to have them notice the frequent use of bundles and various contextual and discoursal functions they perform in academic discipline.

The research article genre was selected because we believed that research articles can be considered as a source of disciplinary knowledge. Without doubt, similar to other texts used in universities, research articles contain lexical bundles which are pervasive in university discourse. Therefore, students encounter these clusters and failure to understand their textual meaning or function leads to failure in their production and comprehension. Consequently,

lack of knowledge in the function of lexical bundles precludes students from the production of these clusters, while the use of lexical bundles leads to the production of fluent spoken discourse and comprehensive and coherent written discourse. Also, research articles are one of the main means by which universities improve and transfer their knowledge and reputation, so identification of lexical bundles grasp specific importance.

Pedagogically, it would be useful if teachers of EAP or EMP courses include lexical bundles in teaching syllabuses as a learning input. They should apply activities which raise awareness toward lexical bundles and show their structures and functions. The output of the current research can help medical researchers in particular to produce more coherent and native-like academic texts.

In the information science field, semantics have been at the centre of attention in retrieval systems. Therefore, in addition to semantics, syntactic structures can play prominent roles in the amelioration of information systems. This research can be considered as an introductory step towards more cooperation among computer, information and linguistics professionals. As a consequence it can help researchers to establish new algorithms for retrieval systems in web search engines and medical scientific databases through using structural and functional analysis of such lexical bundles.

Although this study has investigated the 4-word lexical bundles in all 33 fields in medicine, it would be useful that future studies identify the lexical bundles in each field separately and compare them with each other. Furthermore, another study can be carried out on comparing lexical bundles, their functions and structures in different sections of medical research articles written by Iranian EMP learners and native speakers.

Finally, the linguistic methods have to resolve word ambiguities and/or generate relevant relationships between words. The development of a sophisticated linguistic retrieval system is difficult and it requires complex knowledge bases of semantic information and retrieval heuristics. In addition, these systems often require techniques that are commonly referred to as artificial intelligence or expert systems techniques.

Endnotes

1. http://www.Sciencedirect.com
2. Each medical research article in the medicine and dentistry areas of SDO included 3000 words on average. In order to obtain a corpus of 2420914 from 33 areas, 21 articles were needed from each area separately.

References

Altenberg, B. (1998). On the phraseology of spoken English: the evidence of recurrent word-combinations. in A. P. Cowie (Ed.), *Phraseology: theory, analysis and applications* (pp. 101–122). Oxford: Oxford University Press.

Anthony, L. (2007). Antconc 3.2.1: *A free text analysis software*. Retrieved from http://www.antlab.sci.waseda.ac.jp/

Biber, D. (2004). Lexical bundles in academic speech and writing. in: lewandowska-tomaszczyk B. (Ed.) *Practical Applications in Language and Computers* (pp. 165-178). Frankfurt: Peter Lang.

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

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

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *The Longman Grammar of Spoken and Written English*. London: Longman.

Chen, Y., & Baker, P. (2010). Lexical Bundles in L1 and L2 academic writing. *Language Learning & Technology, 14(2), 30–49.*

Cortes, V. (2002). *Lexical bundles in academic writing in history and biology*. Doctoral dissertation, 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.

De Cock, S. (1998). A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics, 3*(1), 59–80.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. in S. Granger (Ed.), *Learner English on computer* (pp. 67–79). London: Longman.

Dufon, M. (1995). The acquisition of gambits by classroom foreign language learners of Indonesian. in M. Alves (Ed.), *Papers from the 3rd annual meeting of the Southeast Asian Linguistic Society* (pp. 27–42). Tempe: Arizona State University, Program for Southeast Asian Studies.

Erman, B. (2007). Cognitive processes as evidence of the idiom principle. *International Journal of Corpus Linguistics, 12,* 25–53.

Firth, J. R. (1951). Modes of meaning. *Essays and Studies (The English Association),* 118–149.

Gass, s., & Mackey, A. (2002). Frequency effects and second language acquisition. *Studies in Second Language Acquisition, 24,* 249-260.

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning, 24,* 287–297.

House, J. (1996). Developing pragmatic fluency in english as a foreign language. *Studies in Second Language Acquisition, 18, 225–252*.

Hyland, K. (2008a). Academic clusters: text patterning in published and postgraduate writing.

International Journal of Applied Linguistics, 18(1), 1-9.

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

Jalali, H. (2009). *Lexical Bundles in Applied Linguistics: Variations within a Single Discipline.* Unpublished doctoral thesis, University of Isfahan, Isfahan, Iran.

Jespersen, O. (1924). *The philosophy of grammar*. London: George Allen and Unwin.

Karlgren, J. (2000). *Information retrieval; statistics and linguistics. a short introduction to textual information retrieval*. Sweden: Kista, Swedish Institute of computer science, human machine interaction and language engineering laboratory.

Marco, M. J. (2000). Collocational frameworks in medical research papers: a genre based study. *English for Specific Purposes*, 19, 63-86.

Martinez, I. (2003). Aspects of theme in the method and discussion sections of biology journal articles in English. *Journal of English for Academic Purposes, 2*(2), 103–123.

Nattinger, J., & De Carrico, J. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nekrasova, T. (2009). English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning, 59(*3), 647–686.

Parvizi, N. (2011*). Identification of discipline-specific lexical bundles in education.* Unpublished master's thesis, University of Kashan, Kashan, Iran.

Salem, A. (1987). Pratique des segments re´pe´te´s. Paris: Institut National de la Langue Franc¸aise. In Tremblay, A, et al.(2007). Are lexical bundles stored and processed as single units?. *Proc. 23rd Northwest Linguistics Conference, Victoria BC CDA.*

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11, 129–158.*

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic Sequences* (pp. 127–152). Amsterdam: John Benjamins Publishing.

Scott, M. (2008). *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

Stubbs, M. (2007a). An example of frequent English phraseology: distribution, structures and functions. in R. Facchinetti (Ed.), *Corpus Linguistics 25 years on* (pp. 89–105). Amsterdam: Radopi.

Stubbs, M. (2007b). Quantitative data on multi-word sequences in English: the case of word 'world'. in M. Hoey, M. Mahlberg, M. Stubbs & W. Teubert (Eds.), *Text, Discourse and Corpora: Theory and Analysis* (pp. 163–189). London: Continuum.

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

Wang, J., Liang, Sh. & Ge, G. (2008). Establishment of a medical academic word list. *English*

for Specific Purposes 27, 442–458.

Wood, D. (2006). Uses and functions of formulaic sequences in second language speech: an exploration of the foundations of fluency. *Canadian Modern Language Review, 63*, 13–33.

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