sciendo
DOI 10.2478/jazcas-2023-0001

IN-TEXT VARIATION IN THE USE OF LEXICAL BUNDLES IN A CORPUS OF RESEARCH ARTICLES IN PUBLIC HEALTH

TATIANA SZCZYGŁOWSKA
Institute of Neophilology, University of Bielsko-Biala, Poland

SZCZYGŁOWSKA, Tatiana: In-text variation in the use of lexical bundles in a corpus of research articles in public health. Jazykovedný časopis (Journal of Linguistics), 2022, Vol. 73, No. 2, pp. 117–140.

Abstract: This article seeks to contribute to the body of research on the formulaicity of academic written English by adopting a text-internal perspective on the use of lexical bundles across four main rhetorical sections (i.e. Introduction, Method, Results, Discussion) of research articles in public health. The corpus linguistic approach was adopted to explore a section-coded corpus of 200 research papers published in high impact journals with the aim of identifying salient shared and section-specific 4-word bundles, calculating their respective frequencies as well as classifying them structurally and functionally. The results show that the Method is the most formulaic section, containing the greatest number of bundles, including those most frequent and section-specific ones. The sections also differ in the extent to which they exploit the distinct structural and functional categories of shared bundles to fulfill their individual communicative needs.

Key words: academic English, research article, lexical bundles, shared bundles, section-specific bundles

1. INTRODUCTION

Recurrent word combinations that serve specific discourse functions and are familiar to proficient language users constitute manifestations of what is commonly known as "formulaic language" (Wray 2002). The linguistic units that compose it are fixed in form, express a single meaning and behave like single-unit lexical items since they are "prefabricated: that is, stored and retrieved whole from memory at the time of use" (Wray 2002, p. 9). They also help "to shape text meanings", contributing to its naturalness of expression and identifying its register affiliation (Hyland 2008, p. 5). Yet, although such sequences come in many forms (e.g. idioms, proverbs, collocations) and make up around one-half of discourse production (Erman and Warren 2000), they seem to have no single unanimous definition, which is reflected in the plethora of labels attached to them in the literature. Alison Wray (2002, p. 9) listed over fifty such terms, from among which she chose "formulaic sequences" as the umbrella name, though the list is by no means comprehensive, especially that the advancement of corpus linguistics methodology has coined new labels such as "n-grams" (Cheng et al.

2009), "clusters" (Scott 2012) or "lexical bundles" (Biber et al. 1999). In the present study, the focus is on the latter term denoting sequences of three or more words that are frequently repeated in natural discourse and thus constitute basic lexical building blocks of a register or genre (Biber et al. 1999, p. 989–991).

In recent years, lexical bundles (LBs) have become a focal topic in the linguistic analysis of academic discourse, where their fluent use is believed to reflect competent participation in academic community. This belief has not changed much even after Simpson-Vlach and Ellis's (2010, p. 493) observation that many LBs are semantically and structurally incomplete and thus their lists seem "neither terribly functional nor pedagogically compelling". LBs in academic settings have been studied across genres (e.g. Shirazizadeh – Amirfazlian 2021), disciplines (e.g. Durrant 2017) as well as the writers' levels of expertise (e.g. Cortes 2004), first languages (e.g. Lu and Deng 2019), status of (non-)nativeness (e.g. Ädel – Erman 2012), target language proficiency (e.g. Chen 2019) and the English varieties they use (e.g. Liu 2012). Less prolific is research into variation in academic phraseology across the research article sections, particularly in disciplines that are less popular than biology (Cardinali 2015) or linguistics (Shahriari 2017).

The rhetorical structure of what John Swales (2001, p. 50) defines as the "master academic narrative of the last fifty years" has accumulated a body of findings on both the conventional IMRD format of the genre (Swales 1990) and its linguistic exponents (e.g. Kanoksilapatham 2005). One of the important findings is that the distinct sections of empirical research articles are organized in terms of several rhetorical moves and steps realizing particular communicative intentions. Consequently, each section fulfils its own discourse functions, playing its individual role in accomplishing the overall communicative goal of the text. Essentially, the main aim of the Introduction is "to contextualize a research study being presented in the relevant literature, claim its novelty, and present main features of the study" (Kanoksilapatham 2005, p. 274). The Method section presents the research material and provides background information about the study methods, procedures and tools. The Results section is aimed at reporting and commenting on the findings, often with the help of tables and figures. The Discussion section, as Dean R. Hess (2004, pp. 1239–1240) explains, summarizes and interprets the main results, explains their significance and implications, relates them to the findings of previous studies, acknowledges limitations of the study and makes suggestions for further research.

The sections accomplish their discourse aims through various linguistic resources, some of which are section-specific while other ones are more universal. Looking for a connection between those resources and the rhetorical organization of research articles, Swales (1981) compiled a list of words, expressions, and other linguistic features that realize the moves organizing the writers' communicative functions in article Introductions. Viviana Cortes (2013, p. 33) took her analysis a step further, as she considered one specific linguistic feature, namely LBs, and

matched these to the moves and steps found in article Introductions of various disciplines, demonstrating "that a group of lexical bundles were exclusively linked to one move or step in a move while a second group occurred across several moves and steps". In turn, Le and Harrington (2015, p. 55) identified the discourse functions of clusters used to comment on results in Discussions from applied linguistics papers and stated that "clusters in their environment can provide a key to understanding a particular text". What still remains unclear is the extent to which the investigated recurrent word combinations serve as the building blocks of coherent global and text-internal discourse units in the genre of an academic article.

This study attempts to address these issues by examining LBs across four main rhetorical sections (i.e. Introduction, Method, Results, Discussion) of English-language research articles in public health. The aim is to identify salient section-specific and shared bundles, calculate their respective frequencies as well as classify them structurally and functionally. The results obtained will provide a better insight into in-text variation in the use of LBs across the research article sections, thus carrying important implications for novice writers on how to effectively manage academic formulaic language to realize the intended communicative intentions both in the entire text and in its distinct parts.

2. LEXICAL BUNDLES

Douglas Biber et al. (1999, p. 989), who introduced LBs, define them as recurring sequences of three or more words "that show a statistical tendency to co-occur" in natural discourse. The main defining characteristics is a frequency-based method of identification that allows to extract bundles automatically from a given corpus using computer software. The researcher determines the cut-off frequency, depending on the corpus size and research aims, thus setting the frequency threshold at ten (e.g. Biber et al. 1999), through 20 (e.g. Perez-Llantada 2014) up to even 40 instances per million words (Pan et al. 2016). Frequency is also a measure of fixedness, as only a rigidly ordered combination of words that meets the cut-off frequency can be considered as a bundle, irrespective of the existence of variant forms (e.g. *was associated with a*, *was associated with an*). Equally important is the dispersion criterion which prevents user idiosyncrasies, since frequency increases as bundle length decreases (Simpson-Vlach – Ellis 2010), therefore a string must occur in many texts within a register (at least five).

Although LBs usually function as fixed units in discourse, they are semantically transparent rather than idiomatic because their meaning is retrievable from the individual words they contain, as illustrated by such combinations as *do you want to* or *studies have shown that*. Biber et al. (1999) also found that because LBs are identified empirically, not intuitively, only 5% of them in academic prose represent complete structural units and the rest usually cross grammatical boundaries by

bridging two units in such a way that the last word of the bundle serves as the first element of another structural unit. This makes LBs structurally incomplete, but since the majority incorporate fragments of phrases or clauses, they can be classified into basic structural types, such as noun phrase bundles (e.g. *the end of the*), verb phrase bundles (e.g. *were excluded from the*) or dependent clause bundles (e.g. *to be able to*). Additionally, shorter LBs are often subsumed in longer sequences, which according to Chen and Baker (2010, p. 33), can have the form of "complete overlap", where a shorter bundle is derived from a longer bundle and both occur with the same frequency (e.g. *it has been suggested* and *has been suggested that*, both coming from *it has been suggested that*), or "complete subsumption", where two or more "bundles overlap and the occurrences of one of the bundles subsume those of the other overlapping bundle(s)" (e.g. *as a result of* occurs 10 times, while *a result of the* occurs 4 times, and both occur as a subset of *as a result of the*).

The problem of structural and semantic incompleteness of lexical bundles was addressed by Wood and Appel (2014) who proposed a different approach to the identification of LBs. In their first step, they used traditional corpus-driven techniques to generate a list of 4-word sequences that met a frequency cut-off of 25 per million words and a minimum range of 2 university textbooks from among ten that entered the corpus. Subsequently, they refined the obtained list of LBs by producing a list of multi-word constructions of three to five words, identifying 3-word root structures contained within these bundles (e.g. *as long as* as the core structure for *as long as the*) and allowing variable slots for the most likely words typically preceding and/or following each structure (e.g. *(at) the end of (the)*). Owing to this, the researchers obviated the problem of significant overlap among 4-word bundles and could better identify sequences of differing lengths.

Barber and Barbieri (2007, p. 270) note that LBs are not only extremely common but also crucial for the formulation of texts, since "they function as discourse frames for the expression of new information", thus providing "interpretive frames for the developing discourse". For instance, Cortes (2013, p. 41) argues that LBs in introductions to research articles provide structure within clauses, as those occurring at the beginning of a clause start a given rhetorical move by introducing the topic (e.g. *the purpose of the present study*, *little is known about the*), whereas those appearing in the second part of the clause act as "complements" completing the mentioned clause by "adding a commentary" to its main communicative purpose (e.g. *in the sense that*, *can be used to*). Indeed, bundles serve several different functions in discourse, where they assist writers in, for instance, conveying their attitudes (e.g. *it is possible that*), signalling linkage of ideas (e.g. *in addition to the*), or identifying entities (e.g. *the context of the*).

It should be also mentioned that LBs are register-specific, since they differ across genres (e.g. student vs published writing; Cortes 2004), modes (i.e. spoken vs written texts; Biber et al. 2004) and disciplines (e.g. biology vs engineering; Hyland

2008) in both form and function. Therefore, familiarity with these sequences is regarded as a manifestation of users' high linguistic competence and assimilation into a specific discourse community. Yet, contrary to Biber and Conrad's (1999, p. 188) assumption that these ready-made word combinations "are simple expressions" that will "be acquired easily in the natural course of language learning", knowledge of them "is not part of the native writer's innate language ability and is thus far from being a linguistic universal skill" (Pérez-Llantada 2014, p. 84). Indeed, LBs are syntagmatic or collocational associations (Biber et al. 1999), many of which should not be seen as straightforward form-to-meaning mappings that are stored in memory and function like fully predictable templates, on the basis of which systematic generalizations are made to produce and understand novel sequences. LBs have to be learned either through formal instruction or through extensive reading and writing. This makes that studies such as the present one provide valuable insights for all writers of academic English striving to enhance their understanding of how recurrent multi-word sequences contribute to academic prose.

3. CORPUS AND METHOD

3.1. Corpus

The study is based on a corpus of 200 articles in public health (949,800 words) comprising 50 in four leading journals of the discipline: *Environmental Health Perspectives*, *Journal of Global Health*, *Population Health Metrics* and *The Lancet Public Health*. The journals offer open online access to the published papers, all of which have clearly demarcated rhetorical sections required for the study (i.e. Introduction, Method, Results, and Discussion). Each of the journals contributed its 2019 articles, except for *The Lancet* (4 articles from 2018) and *Population Health Metrics* (one article from 2018 and one from 2020), which published fewer articles in the selected year. Only the main body of each article was converted to plain-text format, cleaned by removing notes, long quotations, bibliographies, tables, figures, and subsequently included in the corpus, thus forming four distinct subcorpora, as detailed in Table 1.

Table 1. Corpus composition

Subcorpus	Texts	Words	Average text length
Introduction	200	129,197	615
Method	200	287,878	1,357
Results	200	218,264	967
Discussion	200	312,361	1,510

3.2. Method

Following the approach adopted in previous studies (e.g. Cortes 2004, Hyland 2008, Wright 2019), only the most frequent 4-word bundles were investigated.

Hyland (2008, p. 8) argues that they are more common than 5-word bundles and represent clearer structures and functions than shorter strings. Additionally, "many four-word bundles hold three-word bundles in their structures" (Cortes 2004, p. 401). They also allow to limit the scope of the analysis, as "the number of 4-word bundles is often within a manageable size (around 100) for manual categorization and concordance checks" (Chen and Baker 2010, p. 32).

The bundles were retrieved using WordSmith's (Scott 2012) function WordList cluster. To prevent against idiosyncrasies introduced by particular authors, the frequency threshold was set at 20 occurrences per million words, which was equivalent to a raw frequency of 18 occurrences in the corpus, whereas the cut-off point for range was kept at 5% of the corpus texts, that is, a sequence had to occur in at least 10 different articles. This procedure yielded a total of 219 four-word sequences, which were checked manually by concordancing.

The extraction of the most salient bundles was further facilitated by applying exclusion criteria. The excluded bundles comprised content bundles with proper nouns (e.g. *Centers for Disease Control*, 14 exclusions), as "they are not the 'building blocks' which display a distinct discourse feature" (Chen – Baker 2010, p. 7), and bundles interrupted by a punctuation mark (e.g. *source. The funder of*, 20 exclusions), as suggested by Biber and Barbieri (2007, p. 268). Yet, a bundle was included if only some of its instances were written with a punctuation mark (e.g. *in this study (,) we*). Additionally, following Chen and Baker's (2010, p. 7) procedure guarding against inflated numbers, overlapping bundles with exactly the same frequency, originating from the same contexts, were merged into longer sequences (13 cases), such as *(the)+funder of the study*, which resulted from combining *the funder of the* and *funder of the study*.

After filtering out the data, the final list of 172 target LBs was generated. Based on the taxonomies proposed in previous studies, the target bundles were classified structurally (Biber et al. 1999; Salazar 2014), according to their grammatical types, and functionally (Hyland 2008), according to their contextual meanings and discourse-pragmatic roles. Subsequently, capitalizing on Shahriari's (2017) approach to the study of bundles, which ensures concentration on the most salient strings, the target items were searched for in the four subcorpora to compare their frequency, structure and functions across the research article sections as well as to determine which of them were shared across two or more rhetorical sections and which were section-specific.

4. RESULTS AND DISCUSSION

The following sections discuss the most important findings from the analysis of the use of LBs across different parts of the research article with respect to their frequency, structure and functions.

4.1. Frequency of target bundles

After applying the exclusion criteria, 172 target LBs remained on the list, totalling 5108 individual cases in the overall corpus. Table 2 shows that the Method contained the greatest range of bundles with 149 different items, which constitutes 86.6% of the original list and indicates the section as highly reliant on formulaic sequences. This finding is in line with Cardinali (2015), who analyzed a biology corpus, but contradicts with Shahriari (2017), who focused on linguistics texts and found the Results to be the most densely formulaic part of research papers. Methods also contained the largest number of bundles with a fairly high frequency of more than 40 occurrences per section – 7 items, whereas the remaining three sections contained only one such item each. This may result from the authors' attempt at providing a transparent and replicable description of their research process that, if presented in a routinely patterned way, may be easier to understand for other scholars. The smallest range of bundles occurred in Introductions, where also their frequency was the lowest. This may be partly due to the fact that the Introduction subcorpus is the smallest in terms of word number and average text length. Simultaneously, individual bundles were on average found in every 1153 words of the subcorpus, which was followed by Results with a bundle in every 1865 words, Methods – every 1932 words, and Discussions – every 2263 words.

Table 2. Bundle frequency information

Rhetorical section	***Number of bundles***	***Percentage***	***Frequency***
Introduction	112	65.1	605
Method	149	86.6	1840
Results	117	68	962
Discussion	138	80.2	1701

Considering the distribution of shared and section-specific bundles, summarized in Table 3, it can be seen that the largest percentage of LBs are common to all sections. They thus serve as the building blocks of coherent global discourse of the research articles examined, indicating functional and/or structural similarity in their text-internal linguistic resources. A close similarity is also observed across three sections, particularly across the Introduction, Method and Discussion sections. Bundles found in only two sections are the least numerous, whereas section-specific bundles are more prevalent but were found only in the Method and Discussion. Of these two sections, the former has the highest number of unique bundles that are not found in the other sections. This can be attributed to the fact that 21 of these pre-fabricated structures constitute fragments of a funding acknowledgement statement (e.g. *role of the funding*, *study had no role*, *had final responsibility for*) which is "considered an important element in the scholarly communication process" and is often provided at the end of the Method section (Álvarez-Bornstein et al. 2017, p. 1794).

Table 3. Distribution information on shared and section-specific bundles

Bundles found in … sections											
four	No	%	three	No	%	two	No	%	one	No	%
I, M, R, D	75	43.6	I, M, R	6	3.5	I, D	4	2.3	I	--	--
			I, M, D	20	11.6	I, M	--	--	M	25	14.6
			I, R, D	8	4.6	I, R	--	--	R	--	--
			M, R, D	17	9.9	M, R	2	1.2	D	4	2.3
						M, D	4	2.3			
						R, D	7	4.1			
Totals	***75***	***43.6***		***51***	***29.6***		***17***	***9.9***		***29***	***16.9***

Additionally, two-by-two comparisons revealed that the closest affinity was between Methods and Discussions which shared 116 bundles, although only *were not able to*, *we were not able*, *in our study were* and *in the current study* were exclusive to these two sections. Methods also had 100 bundles in common with Introductions, but all of them also occurred in the other sections, as well as with Results, though only *are presented in Table* was not found in Introductions and Discussions. A greater affinity was observed between Results and Discussions which shared 107 bundles, seven of which were not found in the other sections: *could be due to*, *be interpreted with caution*, *were less likely to*, *we also found that*, *as shown in Figure*, *was associated with an*, *was associated with higher*. Introductions and Discussions had 104 bundles in common, with *low and middle income* being the only bundle exclusive to these two sections. Introductions and Results shared the smallest number of bundles – only 88 – none of which were unique to the two sections.

More specifically, Table 4 shows the *top 35 bundles* ordered by frequency in the entire corpus, as well as their occurrences across the four sections. As can be seen, 23 (66%) are shared by all sections, including the top 10 bundles that all occurred over 30 times per million words and mostly represented seemingly universal strings for academic written English. The exception here are *and middle income countries*, the most common bundle in Introductions (30 instances), and *underlying cause of death*, which are topic-specific. Additionally, *in online supplementary document*, the most frequent bundle with 188 instances in the whole corpus, is typical of *electronic versions of research* papers, where additional material that due to limitations of space or medium cannot be included in the main text is provided in a separate file. Unsurprisingly, it constitutes the most common bundle in Results (91 instances) and Methods (72 instances), both of which often provide large amounts of data or statistics. Table 4 also demonstrates that 10 bundles in the list (28%, italicized) are section-specific, all of which occurred in Methods, whereas only 2 (6%, underlined) are shared by three sections. The list also includes the most common bundle found in Discussions: *in the present study* (41 instances), but it does

not include the most common bundle seen in Introductions: *and middle income countries* (30 instances). Interestingly, only 3 bundles (9%, marked by an asterisk) represent strings that are typical of the topic and the rest are general, that is, typical either of the genre as a whole (shared) or of the distinct sections (section-specific).

Table 4. Top 35 bundles in the corpus and their presence in the four sections

Bundle	***Freq***	Found in			
		Intro	Meth	Res	Disc
in online supplementary document	***188***	6	72	91	19
as well as the	***68***	13	23	10	22
were more likely to	***68***	5	2	38	23
and middle income countries*	***63***	30	1	11	21
at the time of	***62***	2	34	11	15
underlying cause of death*	***59***	12	21	11	15
on the basis of	***58***	6	35	6	11
in this study we	***57***	19	13	1	24
in the present study	***56***	4	7	4	41
it is important to	***52***	9	3	6	34
role of the funding+(source)	***51***		51		
the total number of	***51***	3	27	14	7
more likely to be	***50***	3	6	23	18
to submit for publication	***48***		48		
on the other hand	***48***	8	4	17	19
during the study period	***47***		17	19	11
the number of deaths	***46***	5	14	18	9
had no role in	***45***		45		
had full access to+(all)	***44***		44		
writing of the report	***42***		42		
were included in the	***42***	1	22	18	1
in addition to the	***41***	5	17	5	14
decision to submit for	***40***		40		
has been shown to	***40***	11	3	3	23
for each of the	***39***	1	22	15	1
with the exception of	***39***	3	6	20	10
it is possible that	***38***	3	3	1	31
the corresponding author had	***38***		38		
or writing of the	***38***		38		
the decision to submit	***38***		38		
exposure to air pollution*	***37***	7	2	5	23

(final)+responsibility for the decision+(to)	***37***		37		
was associated with a	***37***	3	1	25	8
more likely to have	***37***	7		22	8
in the control group	***36***	1	4	27	4

Yet a different picture emerges when examining the 35 most commonly used bundles in the four sections. They are presented in Table 5 in frequency order, with items occurring in all four sections marked in bold, those occurring in three sections underlined and those occurring in two sections italicized. It can be seen that over 60% of the items in each list are not shared across the sections, with Methods including as much as 89% (31 items) of such bundles. Only up to 9% of the bundles in each section are found in two other sections and around 30% are found in one other section. Actually, only one item occurring in the top band of bundles is found in all four sections. Generally, the findings in Tables 3, 4 and 5 indicate that notwithstanding the seemingly high number of shared bundles, there are differences in the extent to which they fulfil the individual communicative purposes of a given section, as reflected in their differing frequency of occurrence across the distinct parts of the research article.

Table 5. Most frequent 35 four-word bundles in the sections

Introduction	Method	Results	Discussion
and middle income countries	in online supplementary document	in online supplementary document	in the present study
in this study we	role of the funding+(source)	*were more likely to*	it is possible that
this study was to	to submit for publication	in the control group	this is the first
of this study was	had no role in	was associated with	were not able to
the aim of this	had full access to+(all)	are presented in Table	we were not able
as well as the	writing of the report	*more likely to be*	we were unable to
leading cause of death	decision to submit for	*more likely to have*	*in this study we*
underlying cause of death	the corresponding author had	as shown in Figure	is likely to be
has been shown to	or writing of the	with the exception of	*were more likely to*
have been shown to	the decision to submit	are shown in Table	*has been shown to*
have been associated with		during the study period	*exposure to air pollution*
life expectancy at birth		*were included in the*	**as well as the**
			we found that the
			has been associated with

maternal and child health	(final)+responsibility for the decision+(to)	the number of deaths	and middle income countries
studies have shown that	(access)+to all the data	are shown in Figure	may be due to
of cause of death	(no)+role in study design	on the other hand	could be due to
it is important to	on the basis of	leading causes of death	we were able to
a wide range of	study had no role	*leading cause of death*	*is one of the*
one of the most	at the time of	for each of the	in online supplementary document
an increased risk of	(of)+the study had no	*the total number of*	on the other hand
in the general population	data in the study	less likely to be	*an increased risk of*
cause of death in	(corresponding)+author had full access	was associated with higher	as a result of
the cause of death	(and)+had final responsibility for	the ratio of the	to the best of
as one of the	in the study and	the magnitude of the	*more likely to be*
is one of the	the data in the	in the prevalence of	in the context of
on the other hand	was approved by the	the total number of	*in the number of*
low and middle income	were obtained from the	at the time of	the best of our
the extent to which	all the data in	and middle income countries	it is likely that
more likely to have	*the total number of*	the difference between the	was associated with increased
has been associated with	(the)+study and had final	*in the number of*	loss to follow up
the leading causes of	were excluded from the	the majority of the	*in the general population*
the leading cause of	(the)+funder of the study	*life expectancy at birth*	best of our knowledge
leading causes of death	**as well as the**	*underlying cause of death*	in the absence of
exposure to air pollution	can be found in	was associated with an	of this study is
in low income and	was defined as the	years to follow up	
low income and middle	*were included in the*	**as well as the**	
	for each of the		
	study was approved by		

4.2. Structural analysis of target bundles

The target bundles were classified structurally using Salazar's (2014, p. 68) classification, which is an expanded version of Biber et al.'s (1999, pp. 1014–1024) taxonomy that assigns bundles to four main classes: noun structures, verb structures, prepositional-phrase fragments and other structures. Table 6 presents the main structures of target bundles with their percentage distribution in the corpus and across the rhetorical sections.

Table 6. Structures of bundles across sections

Structure		Intro	Meth	Results	Disc	***Totals***
Noun structures		**34.0**	**32.2**	**31.6**	**28.3**	***29.6***
	Noun phrase + *of*	25.9	23.5	23.9	23.2	*21.5*
	Noun phrase + other modification	6.2	7.4	6.8	3.6	*7*
	Other noun phrase	1.8	1.3	0.9	1.5	*1.1*
Prepositional-phrase fragments		**32.1**	**26.2**	**29.1**	**29.7**	***25.6***
	Prepositional phrase + *of*	16.05	11.4	15.4	13	*11.1*
	Other prepositional phrase	16.05	14.8	13.7	16.7	*14.5*
Verb structures		**12.5**	**24.8**	**21.4**	**19.6**	***25.6***
	Passive + prepositional phrase	8.9	11.4	15.4	11.6	*12.8*
	Other passive fragment	--	0.7	--	--	*0.6*
	Verb phrase with *we*	1.8	4	3.4	5.1	*4.1*
	Other verbal fragment	1.8	8.7	2.6	2.9	*8.1*
Other structures		**21.4**	**16.8**	**17.9**	**22.4**	***19.2***
	Verb/adjective *to*-clause fragment	8.9	8.1	9.4	9.4	***8.1***
	Verb phrase/noun phrase + *that*-clause	1.8	1.3	0.9	1.4	***1.1***
	Adverbial-clause fragment	--	--	0.9	0.7	***0.6***
	Copula *be* + adjective phrase/noun phrase	2.7	2.7	2.5	4.3	***3.5***
	Anticipatory *it* + verb/adjectival phrase	1.8	1.3	1.7	2.2	***1.8***
	Pronoun/noun phrase + *be* (+…)	2.7	2.1	--	2.2	***1.8***
	Other expression	3.5	1.3	2.5	2.2	***2.3***

In line with earlier studies (e.g. Hyland 2008; Salazar 2014; Shahriari 2017), the structures incorporating noun phrases were the most common in the corpus, accounting for almost one third of all bundles. This finding correlates with "the noun-centric style of" English academic writing (Pickering and Byrd 2008, p. 122) and is relevant for all sections except Discussions, where prepositional structures slightly predominated. Noun structures were often used to identify aspects of the topic (*underlying cause of death*, *maternal and child health*), express acknowledgement (*role of the funding+(source)*, *(no)+role in study design*), identify quantity (*a large number of*, *the majority of the*) or provide descriptions (*the quality of the*, *the impact of the*). Also, most of them were variants of the "formulaic frame" *the * of*, where the internal variable slot-filler were various content words, for example, *end*, *cause*, *results*, *time*, *magnitude*, *size* (Biber 2009).

Prepositional-phrase fragments were the second most common structure in Introductions, Methods and Results. The largest group constituted bundles expressing logical relations between propositional elements, all of which had an embedded *of*-phrase (*on the basis of*, *in the case of*, *in the context of*, *in terms of the*, *with the exception of*) and were particularly popular in Introductions. Many prepositional phrases were used for orienting readers through the text (*in the present study*, *in our study we*, *in the current study*, *in online supplementary document*), which was especially noticeable in Discussions. Amounts were also frequently specified through prepositional-phrase fragments – mostly those with *of*-phrases (*in the number of*, *as one of the*) or alternatively, other ones (*more than half of*). Quite a few prepositional-phrase fragments served to identify topic *(in the general population*, *in low income and*) and time (*during the study period, at the time of*).

Verbs structures were the third most common forms in Methods and Results and were the least popular in Introductions and Discussions. Most of them were passive bundles followed by a prepositional phrase fragment, which allowed writers to emphasize "the persuasive strength of data or methodological practice rather than" their own convictions (Hyland – Jiang 2016, p. 270). In line with Salazar's (2014, p. 73–74) claim, passive structures incorporating present-tense verbs often referred to graphic representations of data (*are presented in Table*, *are shown in Figure/ Table*), whereas those with past-tense verbs specified experimental procedures (*were included in the*, *were obtained from the*, *were excluded from the*) and were thus frequently found in Methods and Results. It is also interesting to note that other verbal fragments were the most prominent in Methods, where they were used to acknowledge funding (*the corresponding author had*, *study had no role*). In turn, in Discussions, fairly common were verb phrases with the personal pronoun *we*, which served to discuss observations (*we found that the*, *we also found that*, *we did not observe*), details of the procedure (*we used data from*) and achievements (*we were able to*) or their lack (*we were unable to*, *we were not able*). Hyland (2001) suggests that explicit authorial presence allows to emphasize a writer's contribution and commitment as well as present oneself as a credible researcher, which becomes particularly important when the authors interpret their findings and try to explain them to the readers.

The bundles categorized as other structures were the third most common forms in Introductions and Discussions and were the least popular in Methods and Results. The most common were *to*-clauses preceded by a verb phrase or a predicative adjective, which usually expressed likelihood (*were more likely to*, *is likely to be*, *less likely to be*, *more likely to have*) and sometimes also commented on earlier findings (*has been shown to*, *have been shown to*). Popular were also combinations of the copula *be* followed by an adjective or noun phrase, which expressed causative relations (*may be due to*, *could be due to*, *be due to the*) and allowed to establish a link between the discussed data and other studies (*is in line with*). Interestingly, adverbial-clause

fragments were not seen in Introductions and Methods, which in the case of the former section is not surprising, as such bundles are often used as text-reflexive markers that orient the reader to tables and figures within the text (*as shown in Figure*). Instead, Introductions employed quite a few bundles categorized as other expressions, which can be explained by the fact that most of them denoted aspects of the topic discussed in the article (*low and middle income*, *years of life lost*).

4.3. Functional analysis of target bundles

For functional classification of the target bundles, there was adopted a modified version of Hyland's (2008, p. 13–14) taxonomy[1], which is particularly relevant to academic written genres, as it collects bundles into three research-related groups: research-oriented that "help writers to structure their activities and experiences of the real world"; text-oriented that are "concerned with the organization of the text and its meaning as a message or argument"; participant-oriented that are "focused on the writer or reader of the text". Each of these categories comprises more specific functions, all of which are shown in Table 7 together with their percentage distribution in the corpus and across the rhetorical sections.

Table 7. Functions of bundles across sections

Function		Intro	Meth	Results	Disc	***Totals***
Research-oriented		**44.9**	**39.6**	**44.6**	**38.6**	***36.2***
	Location	3.1	4.7	5	3.5	***3.9***
	Procedure	5.4	7.5	6.5	5.3	***6.3***
	Quantification	12.4	11.1	12.2	11.1	***10.1***
	Description	7	4.7	5.6	5.3	***4.3***
	Topic	17	11.6	15.1	13.4	***11.6***
Text-oriented		**41.7**	**36**	**38.1**	**43.9**	***38.7***
	Transition	3.9	3	3.6	2.3	***2.4***
	Resultative	17.8	14.5	20.9	21.7	***17.9***
	Structuring	6.2	8.1	5	8.2	***7.8***
	Framing	10	8.1	7.9	9.4	***8.2***
	Citation*	3.8	2.3	0.7	2.3	***2.4***
Participant-oriented		**13.2**	**24.4**	**17.3**	**17.5**	***25.1***
	Stance	10.6	10.5	15.8	16.4	***13.5***
	Engagement	1.6	1.1	1.5	1.1	***1***
	Acknowledgement*	--	12.8	--	--	***10.6***

The analysis showed that 35 bundles were multifunctional, which was taken into consideration when collecting data on their distribution in the corpus.

[1] The subcategories that were added to Hyland's taxonomy, marked with an asterisk, were borrowed from Salazar (2014, p. 91)

Specifically, when a bundle was identified as performing multiple functions in different contexts, as illustrated by *has been shown to*, which in (1) serves as a resultative signal and in (2) helps to refer to prior research, it was included in both functional categories. Additionally, once a multifunctional bundle was found in one of the rhetorical sections, its instances of use were carefully analyzed to see how many cases should be considered under which functional category. Overall, 172 target bundles served 207 different functions, from among which 172 were identified in Methods, 171 in Discussions, 139 in Results and 129 in Introductions. These numbers reflect the results obtained for bundle frequency across the rhetorical sections, which was the highest in Methods, and the lowest in Introductions (see Table 2).

(1) *Mortality from ischemic heart disease has been shown to vary by source of PM2.* [Discussion]
(2) *[…] TPHP has been shown to promote adipogenesis in human primary preadipocytes (Tung et al. 2017).* [Discussion]

It can be seen from Table 7 that research-oriented and text-oriented bundles were used with a comparable frequency in the entire corpus, whereas participant strings were less frequent. This is consistent with the findings of Hyland (2008, p. 14) and Salazar (2014, p. 91), similarly as the prevalence of text-oriented bundles over other functional categories that was observed in the latter study but not in the former, where research-oriented bundles predominated. It has to be noted, however, that Hyland derived his data from the corpora of applied, pure and social sciences, whereas Salazar, from texts in the health sciences (i.e. biology, biochemistry, biomedicine and medicine), which are partly related to the corpus examined here. The overall distribution trend of target bundles was reflected only in Discussions, since in the other sections, the most frequent were research-oriented bundles. The more specific discourse functions of these strings will be discussed in the next three subsections.

4.3.1. Research-oriented bundles

One clear difference between the rhetorical sections is the preference for research-oriented bundles particularly in Introductions and Results, but also in Methods, which emphasizes the role they play in "giving an objective, unbiased and precise account of experimental procedures, so that the subsequent data interpretation can be established as verifiable, reproducible and grounded in empirical reality" (Salazar 2014, p. 97). In all four sections, the majority of such strings introduced aspects of the topic and were, therefore, particularly common in Introductions, which serve to present readers with the subject matter and the context of the research. These bundles were mainly realized by noun phrases (3) and definitely less frequently by prepositional phrases (4).

(3) *Exposure to air pollution and road traffic noise, for example, have both been associated with increased risks of diabetes and hypertension […]* [Intro]
(4) *Madagascar is no exception; its CRVS system is still inoperative at the national level, […]* [Intro]

Several prepositional (5) and noun phrases (6), usually the ones comprising *of*, also denoted quantification and were fairly evenly distributed across all rhetorical sections.

(5) *The increase in recent decades in the proportion of births occurring in health facilities in sub-Saharan Africa has been remarkable.* [Disc]
(6) *The ratio of the travel time by car/public transport was converted into four categories […]* [Meth]

Bundles specifying research procedures and showing how the study was conducted were the third most common subtype of research-oriented strings in Methods and Results. This is not surprising, as both sections are closely focused on methodological and analytical considerations and serve as a venue for the presentation of experimental data rather than for the interpretation of these phenomena. These bundles were mainly realized by past-tense passive structures followed by a prepositional phrase (7), though instances of other realization patters were also seen in the corpus (8).

(7) *Cases missing serotype data were included in the analyses of all serovars but not included in the analyses of the environmental serovars.* [Meth]
(8) *We used data from the household survey to construct a social capital index and data from the entomological survey to calculate vector indices for each household.* [Meth]

Bundles providing descriptions were the most popular in Introductions, where they typically outlined the background to the specific study (9), similarly as in Discussions (10), but unlike in Methods and Results, where they often described research objects, materials, equipment and models (11).

(9) *Although early menopause has been linked to an increased risk of cardiovascular disease mortality and all-cause mortality, few studies have examined […]* [Intro]
(10) *In the previous literature, the economic perspective has been frequently adopted for the evaluation of the impact of the HPV-related diseases.* [Disc]
(11) *The size of the exposed population in each age category (di) was assumed to be […]* [Meth]

Location bundles were the least common in the four sections, where the majority indicated time and were usually realized either by noun (12) or prepositional phrases (13) with *of*. Their larger concentration in Methods and Results can be attributed to the fact that they often indicated temporal details involved in the methodological framework.

(12) *Three senior research officers supervised the field data collection team who received intensive one-week training before the start of the study.* [Meth]
(13) *[…], which was in direct contact with aerosolized uranium particles at the time of the inhalation procedures.* [res]

4.3.2. Text-oriented bundles

Text-oriented bundles were the most common in Discussions, which reflects the more interpretative and evaluative but simultaneously less empiricist nature of the section. In this part of the research article, the writer's aim is to explain the significance of their findings, also by relating them to the established literature, as well as to guide readers both around the text and towards the desired conclusions, which requires a more discursively elaborate style of writing.

In line with previous studies exploring biology-related writing (Hyland 2008; Salazar 2014) – both in the entire corpus and across the rhetorical sections – there was a preponderance of resultative markers indicating "inferential or causative relations between elements" (14), highlighting conclusions (15) and developing interpretations of the results (16) (Hyland 2008, p. 14). These functions account for a higher concentration of such bundles in Results and Discussions which revolve around the presentation and interpretation of findings. The majority of resultative bundles incorporated a verb form, which, as the examples show, was either used in the passive, combined with the pronoun *we* or with a *to*-clause fragment.

(14) *As low cardiorespiratory fitness has been associated with at least 26 chronic diseases, […]* [Disc]
(15) *We also found that long-term exposures to NO2 and O3 were positively (albeit marginally) associated with AF.* [Disc]
(16) *This discrepancy might be due to the cross-sectional study design or the relatively short study period.* [Disc]

The second most common type of text-oriented strings were framing devices contributing to "the effective elaboration of arguments" (Salazar 2014, p. 101), usually in the form of preposition + *of* structures, which helped writers to define conditions (17), highlight connections (18), define limitations (19) and specify cases (20). They were fairly evenly distributed across the rhetorical sections, but were the most common in Introductions, where they assisted writers in providing a plausible background for the study.

(17) *[…] this is also an indicator of household SES within this population, especially in the absence of individual-level or household-level measures of income.* [Disc]
(18) *Second, we explore the implications of the ageing of AIDS in terms of the increasing role of NCDs among HIV-affected individuals, […]* [Intro]
(19) *The intervention was fully implemented in Group 1 schools at visit 4, with the exception of Samoui district, […]* [Disc]
(20) *In the case of deaths certified after inquest, the coroner supplies the registrar with all the particulars that would have been supplied by the informant; […]* [Meth]

The next most frequent subtype of text-oriented bundles were structuring devices, which in Methods were as common as framing bundles. Structuring signals dominated in Methods and Discussions, where they served to organize the text (21), also by referring readers to tables or figures (22), introduce research aims (23) or limitations (24) and provide a frame for the presentation of arguments (25). It is clear from the examples that these functions are performed by a variety of structural patterns, all of which contribute to the creation of a coherent and well-arranged paper.

(21) *More details on the survey data used in our analyses can be found in the appendix (p 3).* [Meth]
(22) *All of the active structures for this class are shown in Figure S2B.* [Res]
(23) *The aim of this study was to analyze the epidemiological trends of hospital admissions, […]* [Intro]
(24) *The primary limitation of this study is the nonspecificity of DAPs.* [Disc]
(25) *In this study we assess whether levels of retention in care and on treatment have changed over time […]* [Disc]

Differences between the sections were observed regarding transition and citation bundles. In Methods and Results, transitions prevailed over citations, whereas in Introductions and Discussions, both types of strings were used with almost the same frequency. Considering Swales (2004, p. 234) claim that "discussions tend to offer a mirror-image reversal of the move-order in Introductions", it is not surprising that both sections provide contrastive (26) or additive (27) links between propositional elements equally often as they refer to topic-specific literature (28). Also, while transitions tended to take the form of prepositional phrases, citations were often realized by verb structures.

(26) *On the one hand, disaggregated data are needed to obtain more accurate results in spatial data analysis. On the other hand, lower disaggregation tends*

to better preserve individual's confidentiality and contribute to more robust rates. [Intro]

(27) *In addition to the homeostasis of cholesterol, arsenic also influences lipid and fatty acid metabolism.* [Disc]

(28) *The lack of an association with road traffic noise is in line with a recent meta-analysis (Fuks et al. 2017).* [Disc]

4.3.3. Participant-oriented bundles

Participant-oriented bundles were the least common both in the entire corpus and across the rhetorical sections, where their frequency was the highest in Methods and the lowest in Introductions. As Salazar explains (2014, p. 103), participant-oriented bundles correspond "to the dialogic interaction between the participants in the text: the writer and the reader" and provide "a structure for interpreting a following proposition" (Hyland 2008, p. 18).

As can be seen from Table 7, many participant-oriented strings served writers to convey stance, that is, "epistemic and affective judgements, opinions and degrees of commitment to what they say", owing to which an authorial presence can be construed in texts (Hyland 2009, p. 111). Stance markers expressed a variety of meanings, most of which were concerned with hedging the effect of assertions by incorporating some dose of tentativeness (29), probability (30) or possibility (31), and only some conveyed stronger meanings, such as necessity (32). The marked presence of such bundles in biology-related writing was also reported by Cortes (2004) and Salazar (2014). As the examples show, stance expressions were realized by diverse structural patterns, usually either those comprising the copula *be* or a *to*-clause fragments, but also anticipatory *it* constructions. An important finding is that the majority of stance bundles in the corpus were impersonal, indicating the writers' desire to present their claims as plausible opinions rather than accredited facts. According to Swales (1990, p. 175), by "projecting honesty, modesty and proper caution in self-reports", as reflected in the increased frequency of stance bundles in Results, and "diplomatically creating space in areas heavily populated by other researchers", as reflected by the high frequency of stance bundles in Discussions, writers can report and interpret their research results in an objective and yet precise way.

(29) *Therefore, the observed results may be due to a lifetime altered methyl pool availability and SAM, […]* [Disc]

(30) *Also, because lower socioeconomic groups might be more likely to have poorer diets and higher proportions of disease attributable to diet, […]* [Disc]

(31) *Thus, it is possible that pollutants with LXR-antagonist activities may induce atherosclerosis by promoting foam cell formation.* [Intro]

(32) *Therefore, there is a need for expanded and improved risk factor surveillance systems, including unrecorded alcohol use in particular.* [Disc]

Another interesting issue that emerged from the data was the link between stance bundles and text-oriented resultative bundles, which was also observed by Salazar (2014, p. 105). These two functions were performed, for instance, by bundles comprising the copula *be* (33) or a *to*-clause fragment (34). Yet, some of such bundles incorporated the pronoun *we* (35), which allowed writers to explicitly signal their responsibility for the claims made. As the examples show, the drawing of inferences from empirical findings combines well with the expression of "personal attitudes and assessments of the status of knowledge" (Hyland – Jiang 2016, p. 253).

(33) *This discrepancy might be due to the cross-sectional study design or the relatively short study period.* [Disc]
(34) *Moreover, Syrian children less than five years of age were significantly less likely to be fully immunized than Afghani children less than five […]* [Res]
(35) *Because household identifiers were no longer collected after 2014, we were unable to account for clustering of children at the household level.* [Meth]

The above mentioned concentration of stance bundles was the least prominent in Methods, which as the only rhetorical section was dominated by acknowledgement bundles explaining the role of the funding source (36) and the respective duties of the authors (37). These meaning were usually conveyed by verbal fragments.

(36) *The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report.* [Meth]
(37) *The authors had full access to all the data in the study and the corresponding author had final responsibility for the decision to submit for publication.* [Meth]

The least common subtype of participant-oriented bundles were those marking engagement with readers by involving them "as participants in the unfolding discourse" (Hyland 2008, p. 18). Engagement markers prompt readers to focus on specific details that are thus seen from the perspective suggested by the writer, which, as Salazar (2014, p. 106) argues, convinces them "to adopt the writer's position". In fact, there were only two such bundles in the entire corpus *it is important to* and *important to note that*, both of which were part of a longer bundle formulated as a directive: *it is also important to note that* (38). Engagement strings were most common in Introductions, possibly due to the writers' attempt at directing readers towards the intended meanings from the very beginning of the research paper.

(38) *It is important to note that the presence of biotechnologies was strong under the umbrella term of novel technologies, referring primarily to nanotechnologies.* [Res]

4.4. Additional comments on section-specific bundles

As mentioned above, the majority of section-specific bundles found in Methods (i.e. 21) were participant-oriented, functioning to provide information about founding. Among the remaining ones, three were research-oriented and, more specifically, concerned with the details of research procedures (*was approved by the*, *were obtained from the*, *study was approved by*), whereas one bundle was text-oriented and used as a framing device (*was defined as the*). Regarding the structural correlates of LBs in Methods, verb structures were the most frequent (14 items), followed by noun structures (8 items), prepositional phrases (2 items) and other structures (1 item).

Regarding the four section-specific bundles found in Discussions, three were text-oriented and could be further sub-categorized as framing (*(to)+our knowledge this is+(the)*), resultative (*we did not observe*) and structuring (*in our study we*) devices. One bundle (i.e. *it is likely that*) was multifunctional, as it simultaneously functioned as a participant-oriented stance and text-oriented resultative device. Looking at these bundles, it can be seen that the first two are verb structures, the third is a prepositional-phrase fragment and the fourth is the *anticipatory it structure.*

5. CONCLUSION

The main aim of the study was to examine the frequency, structure, and function of salient 4-word bundles found in the IMRD sections of research articles in public health. The results reveal something about the extent to which the discussed patterns of formulaicity contribute to the construction of a coherent and integrated academic text, *on the one hand*, and, *on the other*, emphasize the discursive autonomy of each specific section.

Regarding frequency, only less than a quarter of bundles were section-specific and the rest were shared by all sections. However, the prominence of the latter strings varied across the sections, depending on the sections' main functional aims. This finding confirms Shahriari's (2017) conclusion drawn for the discipline of applied linguistics that the discourse functions of LBs differ across particular sections of academic papers. The Method proved to be the most formulaic section, with the largest number of different LBs, including those most frequent and also section-specific ones. As for the main structures, Discussions were dominated by prepositional-phrase fragments, whereas the other sections, by noun structures, particularly those followed by *of*. Yet, only Methods and Results, and to a lesser degree, also Discussions included a decent number of verb structures, among which the most common were passive bundles followed by a prepositional fragment. Regarding the main functions, Introductions, Results and partly also Methods contained the greatest number of research-oriented bundles, particularly those relating to aspects of the topic and quantification. In Discussions, the most prominent

were text-oriented bundles, especially those marking resultative relations, which were also common in the other sections. Participant-oriented bundles were moderately frequent in the sections, but quite a few of them were found in Methods, where the subcategory of acknowledgement, providing information about founding, accounted for almost all section-specific bundles in this part of the article. The other sections were dominated by stance bundles.

Despite the fact that the obtained results concern only one discipline and may involve some bias related to their choice for the final list and the accurate recognition of their discourse functions, they have potential implications for novice and experienced academic writers. Essentially, the study raises awareness of the most salient LBs in different research article sections, revealing the similar and dissimilar ways in which these strings occur and behave in each specific part of the article. Familiarity with the tendencies in the use of formulaic expressions can help writers to improve the pace and quality of their efforts to communicate research findings accurately, effectively and engagingly to scientific and general audiences alike.

References

ÄDEL, Annelie – ERMAN, Britt (2012): Recurrent word combinations in academic writing by nativeand non-native speakers of English: A lexical bundles approach. In: *English for Specific Purposes*, Vol. 31, pp. 81–92.

ÁLVAREZ-BORNSTEIN, Belén – MORILLO, Fernanda – BORDONS, María (2017): Funding acknowledgments in the Web of Science: completeness and accuracy of collected data. In: *Scientometrics*, Vol. 112, pp. 1793–1812.

BIBER, Douglas – CONRAD, Susan (1999): Lexical bundles in conversation and academic prose. In: H. Hasselgard – S. Oksefjell (eds.): *Out of corpora: Studies in honor of Stig Johansson*. Amsterdam: Rodopi, pp. 181–189.

BIBER, Douglas – JOHANSSON, Stig – LEECH, Geoffrey – CONRAD, Susan – FINEGAN, Edward (1999): *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education Limited. 1204 p.

BIBER, Douglas – CONRAD, Susan – CORTES, Viviana (2004): If you look at …: Lexical bundles in university teaching and textbooks. In: *Applied Linguistics*, Vol. 25, No. 3, pp. 371–405.

BIBER, Douglas – BARBIERI, Federica (2007): Lexical bundles in university spoken and written registers. In: *English for Specific Purposes*, Vol. 26, pp. 263–286.

BIBER, Douglas (2009): A corpus-driven approach to formulaic language in English: multi-word patterns in speech and writing. In: *International Journal of Corpus Linguistics*, Vol. 14, No. 3, pp. 275–311.

CARDINALI, Renata F. (2015): Lexical bundles in biology research articles: structure and function across corpora and sections. Available at: https://www.researchgate.net/publication/276276215_Lexical_bundles_in_biology_research_articles_structure_and_function_across_corpora_and_sections [cit. 24.05.2021].

CHEN, Alvin Cheng-Hsien (2019): Assessing phraseological development in word sequences of variable lengths in second language texts using directional association measures. In: *Language Learning*, Vol. 69, No. 2, pp. 440–477.

CHEN, Yu-Hua – BAKER, Paul (2010): Lexical bundles in L1 and L2 academic writing. In: *Language Learning and Technology*, Vol. 14, No. 2, pp. 30–49.

CHEN, Yu-Hua – BAKER, Paul (2014): Investigating Criterial Discourse Features across Second Language Development: Lexical Bundles in Rated Learner Essays, CEFR B1, B2 and C1. In: *Applied Linguistics*, Vol. 37, No. 6, pp. 1–33.

CHENG, Winnie – GREAVES, Chris – SINCLAIR, John McH. – WARREN, Martin (2009): Uncovering the extent of the Phraseological tendency: Towards a systematic analysis of concgrams. In: *Applied Linguistics*, Vol. 30, pp. 236–252.

CORTES, Viviana (2004): Lexical bundles in published and student disciplinary writing: Examples from history and biology. In: *English for Specific Purposes*, Vol. 23, No. 4, pp. 397–423.

CORTES, Viviana (2013): The purpose of this study is to: Connecting lexical bundles and moves in research article introductions. In: *Journal of English for Academic Purposes*, Vol. 12, pp. 33–43.

DURRANT, Philip (2017): Lexical bundles and disciplinary variation in university students' writing: Mapping the territories. In: *Applied Linguistics*, Vol. 38, No. 2, pp. 165–193.

ERMAN, Britt – WARREN, Beatrice (2000): The idiom principle and the open-choice principle. In: *Text*, Vol. 20, pp. 29–62.

HESS, Dean R. (2004): How to Write an Effective Discussion. In: *Respiratory Care*, Vol. 49, No. 10, pp. 1238–1241.

HYLAND, Ken (2001): Humble servants of the discipline? Self-mention in research articles. In: *English for Specific Purposes*, Vol. 20, No. 3, pp. 207–226.

HYLAND, Ken (2008): As can be seen: Lexical bundles and disciplinary variation. In: *English for Specific Purposes*, Vol. 2, No. 1, pp. 4–21.

HYLAND, Ken (2009): Corpus informed discourse analysis: The case of academic engagement. In: M. Charles – D. Pecorari – S. Hunston (eds.): *Academic writing: At the interface of corpus and discourse*. London: Continuum, pp. 110–128.

HYLAND, Ken – JIANG, Kevin (2016): Change of attitude? A diachronic study of stance. In: *Written Communication*, Vol. 33, No. 3, pp. 251–274.

KANOKSILAPATHAM, Budsaba (2005): Rhetorical structure of biochemistry research articles. In: *English for Specific Purposes*, Vol. 24, No. 3, pp. 269–292.

LE, Thi Ngoc Phuong – HARRINGTON, Michael (2015): Phraseology used to comment on results in the Discussion section of applied linguistics quantitative research articles. In: *English for Specific Purposes*, Vol. 39, pp. 45–61.

LIU, Dilin (2012): The most frequently-used multi-word constructions in academic written English: a multi-corpus study. In: *English for Specific Purposes*, Vol. 31, pp. 25–35.

LU, Xiaofei – DENG, Jinlei (2019): With the rapid development: A contrastive analysis of lexical bundles in dissertation abstracts by Chinese and L1 English doctoral students. In: *Journal of English for Academic Purposes*, Vol. 39, pp. 21–36.

PAN, Fan – REPPEN, Randi – BIBER, Douglas (2016): Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. In: *Journal of English for Academic Purposes*, Vol. 21, pp. 60–71.

PÉREZ-LLANTADA, Carmen (2014): Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. In: *English for Specific Purposes*, Vol. 14, pp. 84–94.

PICKERING, Lucy – BYRD, Pat (2008): Investigating connections between spoken and written academic English: Lexical bundles in the AWL and in MICASE. In: D. Belcher – A. Hirvela (eds.): *The Oral/Literate Connection: Perspectives on L2 Speaking, Writing and Other Media Interactions*. Ann Arbor MI: University of Michigan Press, pp. 181–189.

SALAZAR, Danica (2014): *Lexical Bundles in Native and Non-native Scientific Writing. Applying a corpus-based study to language teaching*. Amsterdam/Philadelphia: John Benjamins. 212 p.

SCOTT, Mike (2012): *WordSmith Tools* (Version 6). Stroud: Lexical Analysis Software.

SHAHRIARI, Hesamoddin (2017): Comparing lexical bundles across the introduction, method and results sections of the research article. In: *Corpora*, Vol. 12, No. 1, pp. 1–22.

SHIRAZIZADEH, Mohsen – AMIRFAZLIAN, Rojan (2021): Lexical bundles in theses, articles and textbooks of applied linguistics: Investigating intradisciplinary uniformity and variation. In: *Journal of English for Academic Purposes*, Vol. 49, pp. 1–13.

SIMPSON-VLACH, Rita – ELLIS, Nick C. (2010): An academic formulas list: new methods in phraseology research. In: *Applied Linguistics*, Vol. 31, No. 4, pp. 487–512.

SWALES, John (1981): *Aspects of article introductions*. Birmingham, UK: The University of Aston, Language Studies Unit. 104 p.

SWALES, John (1990): *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press. 260 p.

SWALES, John (2001): EAP-related linguistic research: An intellectual history. In: J. Flowerdew – M. Peacock (eds.): *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press, pp. 42–54.

SWALES, John (2004): *Research genres: Exploration and applications*. Cambridge: Cambridge University Press. 314 p.

WOOD, David C. – APPEL, Randy (2014): Multiword constructions in first year business and engineering university textbooks and EAP textbooks. In: *Journal of English for Academic Purposes*, No. 15, pp. 1–13.

WRAY, Alison (2002): *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press. 332 p.

WRIGHT, Heidi R. (2019): Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. In: *English for Specific Purposes*, Vol. 54, pp. 1–14.