Hacettepe Üniversitesi Eğitim Fakültesi Dergisi

Hacettepe University Journal of Education

e-ISSN: 2536-4758

Corpus Analysis of the Use of Lexical Bundles by Turkish EFL Learners at a Tertiary Level*

Hakan TAŞKAYA, Ali Şükrü ÖZBAY****

Article Information

Received: 25.12.2021

Accepted: 30.05.2023

Online First: 21.07.2023

Published: 31.07.2023

ABSTRACT

This study investigated lexical bundle overuse-underuse patterns of English as a Foreign Language (EFL) learners. The functional analysis identified the frequencies, overuse-underuse patterns of the lexical bundles from L2 English expository argumentative and academic essays. The analysis was done with the first most frequent 100 lexical bundles. The cut-off point of frequency was ten times per million. The normalized frequencies of each bundle were tabled, and their log-likelihood scores were calculated and given in the subsequent tables. Functional analysis was done in all categories. Findings indicated not only homogeneous use in academic and expository texts, but also non-native learners used a restricted number of bundles, and finally, underuse and overuse patterns were noted. In other words, EFL learners avoided using common bundles and their usage patterns were limited to a few but repeated patterns. Explicit teaching of the most common lexical bundles should be integrated into the curriculum and writing instructors and material writers (developers) should emphasize these word groups in courses and coursebooks.

Keywords: Corpora, word patterns, functional categories, academic, argumentation, EFL

doi: 10.16986/HUJE.2023.494 Article Type: Research Article

Citation Information: Taşkaya, H., & Özbay, A. Ş. (2023). Corpus analysis of the use of lexical bundles by Turkish EFL learners at a tertiary level. *Hacettepe University Journal of Education, 38*(3), 335-346. doi: 10.16986/HUJE.2023.494

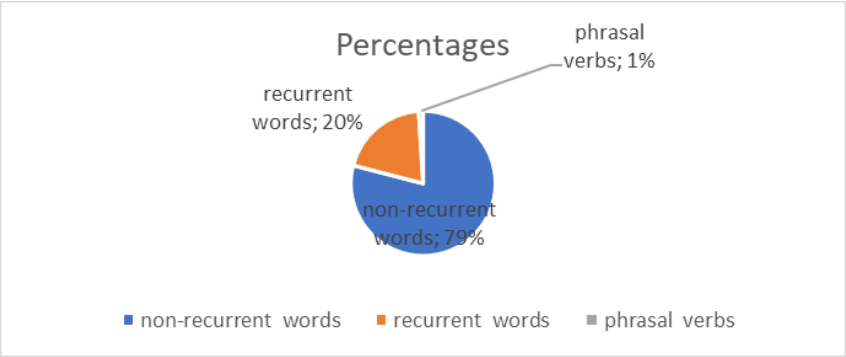
1. INTRODUCTION

Lexical bundles are word combinations mostly occurring in sentences which are "important building blocks of discourse" and standing as functional units in written and spoken contexts, according to Biber and Barbieri (2007, p. 263). Rather than composing a new term, lexical bundles refer to the first meaning of the words in a general sense and act as recurrent expressions without idiomaticity and structure. Wei & Lei (2011, p. 156) points out that "lexical bundles, fixed expressions and collocations are considered to play an important role in both oral and written language production and language learning". Being sequences of word forms used in natural discourse, lexical bundles do not involve much flexibility in themselves when compared to idiomaticity and collocations. Used both in speech and writing, lexical bundles do not contain fixed meanings and differ from idioms in several ways. For example, while their functions and the meaning are understood through the analysis of the individual words, this is not the case if we are working with idioms with no context and contextual data. According to many researchers such as Biber et al. (2004), Biber, Conrad and Cortes (2003), and Cortes (2004), these word combinations are largely observed in academic prose and carry important discursive functions in addition to being structural correlates with their distinctive features in academic prose. Being a sub-group of these phraseological structures, Biber et al. (1999, p. 990) describe these recurrent combinations as a "group of words that show a statistical tendency to co-occur as well as recurrent expressions regardless of their idiomaticity and structural status". They are identified empirically in a given register thanks to the corpus analysis tools. Biber et al. (1999) also state that the bundle variety is greater in academic prose, adding that they are used as part of noun and prepositional phrases in academic prose. They are also seen as extended collocations consisting of three, four or five words. They carry various other meanings based on their functions, but rather remain as lexical units which are functionally effective in written and spoken contexts (Biber et al.,1999; Hyland, 2008a). Wei and Lei (2011, p. 37) state that "Lexical bundles contrast with idioms, which are the whole phrases with meaning unrelated to the parts". The fact that "lexical bundles are crucial for constructing a discourse in university register" makes it clear that having formulaic language competency is likely to foster the students' ability to use the language more naturally (Biber, 2006, p. 74). "Conversely, misuse of formulaic language has been shown to be a potential source of communication difficulty" (Miller, 2009, p. 13) and that "lexical bundles take crucial roles in academic writing and conversations and constitute nearly 20 % of the academic prose while they are used 28 % in conversation" (Biber, 1999, p. 989).

* This study is partly based on the first author's MA Thesis.

** Lecturer, Artvin Çoruh University, International Relations Office, Artvin-TÜRKİYE. e-mail: hakantaskaya@artvin.edu.tr (ORCID: 0000-0001-9084-8715)

*** Assoc. Prof. Dr., Karadeniz Technical University, Faculty of Literature, Department of Western Languages and Literature, Trabzon- TÜRKİYE. e-mail: ozbay@ktu.edu.tr (ORCID: 0000-0002-3421-0650)



Graphic 1. Recurrent and non-recurrent words in academic prose (Biber et al. 1999, p. 994)

In another study, Altenberg (1998) stated that recurrent sequences are largely used in a language (80%). This view is supported by other researchers such as Pawley and Syder (1983), Sinclair (1991), Cowie (1998) and Wray and Perkins (2000), who noted that many languages largely contain lexical properties which are repetitive in nature. Howarth (1998), Granger (1998), and Erman (2009) also stated that L2 learners use these bundles in restricted numbers while the phenomenon has a big usage potential in L1 speech.

Briefly, our study indicates that Turkish EFL writers used few bundle patterns frequently and the diversity of them seemed to be restricted to a few common bundles compared to the research findings in other settings.

1.1. Statement of the Problem

Lexical bundles are significant components of fluency in the language, and language learners need to have good control over them by developing sensitivity to native speakers' preferences (Hyland, 2008a). Hoey (2000, p.202) clearly states this need by arguing that "learning is not just the meanings of the words but the environments in which they occur". The identification of these environments may help learners differentiate various rhetorical functions of texts they need to produce in terms of lexical properties (Chang & Kuo, 2011). Regarding this, Chen and Baker (2010, p. 34) noted that "L1 learners use formulaic language more than L2 learners", and Turkish EFL learners are not an exception. We argue that more studies related to formulaic language use may contribute to efforts to foster language teachers, and course designers initiate attempts to bridge this gap. The study of Öztürk and Köse (2016) in the Turkish context also shows that there are overuse patterns in the ways the bundle patterns are employed by Turkish EFL learners from variety and frequency dimensions, which is also confirmed by Bal (2010) and Güngör and Uysal (2016), who underlined the fact that Turkish EFL learners employ many lexical bundles. However, it is seen by comparison that most of their preferences are with some frequently overused items, not seen in L1 writers' texts. This signals the problem that tertiary EFL learners experience with bundles. The main problem is, according to Cortes (2004), that there is a failure in using them properly compared to L1 writers, and 'repetitions' are constant in L2 writing productions. We understand that L2 learners need to use these bundle groups, which are part of disciplinary conventions, and which are largely used by academics (Hyland, 2008a) with an informed attitude. To date, few studies have investigated the role and importance of learning lexical bundles for EFL learners in argumentative writing in a Turkish context. In the following section, the aim and the methodology are given with a focus on the lexical bundle types, frequencies and the corpora used for investigation.

1.2. Purpose of the Study

From a broader perspective, Biber (2009), stressing its importance for language learners, emphasizes the view that phraseological patterns or formulaic language constitute a significant part of conversational and written academic discourse, adding that fixed sequences representing clause fragments in speech as opposed to formulaic frames consisting of noun and prepositional phrase fragments are the important components of the English lexicon. This may indicate a need for tertiary level EFL learners to learn these frames or word sequences with informed attention. Sinclair (2004, p. 29) also underlined the fact that "phraseological tendency occurs in a language where new meanings are created through word combinations". Parallel to the increasing importance and gaining a critical position in learning and teaching, several studies were made on the lexical properties of texts, focusing on the use of authentic language by EFL writers in various levels and settings. Accordingly, our investigation aimed to analyze the most frequent bundles in six corpora that included data by native and non-native learners, including Turkish EFL learners (KTUCALE and TICLE) from a functional perspective based on the theoretical framework of contrastive interlanguage analysis (CIA) (Granger, 1998).

2. METHODOLOGY

In all phases of this study, research and publication ethics are complied with. We followed Granger's (2009) Contrastive Interlanguage Analysis (CIA) approach in the study.

The research questions that were answered in the study:

1. What characterizes tertiary level Turkish EFL learners' lexical bundles awareness in prose?
2. What are the differences, if any, in the bundle performance of native and non-native writers?
3. What are the preferred lexical bundles in a corpus of argumentative and academic essays by Turkish EFL learners?
4. What are the lexical bundle use performance differences of in the essays of Turkish EFL learners?
5. How are the bundle patterns classified on the basis of functional characteristics?

In this section, we made a lexical bundles corpus analysis and indicated the methodology, the data, and the analytical framework. Argumentative and academic essays of tertiary-level EFL students were analyzed through corpus-based contrastive interlanguage analysis. We also aimed to investigate the lexical bundle usage patterns and the corpus-based contrastive analysis included six corpora. The corpora used in the study consisted of texts related to expository and academic argumentation to a large extent. Table 1 below shows the corpora used in the study.

Table 1.
Corpora Used in the Study

Corpus	Tokens	Texts	Native/ Non-Native	Expository/ Academic
KTUCALE	819846	220	Non-Native	Academic argumentation
BAWE	624294	223	Native	Academic argumentation
TICLE	223449	287	Non- Native	Expository argumentation
LOCNESS	361054	372	Native	Expository argumentation
AUGER	2300000	Not available	Native	Academic texts
LSWE	40000000	Not available	Native	Academic texts

The corpora listed in the above table represent general writing skills in different cultures. This representativeness helps us to make a generalization related to the population of learners. The current study used frequency-based analysis for the most part. There are five steps in the current study.

Table 2.
Study Design

Step 1	Specification of target corpora
Step 2	Data Collection: Criteria for the selection was given and the list of the bundles based on Sketch Engine.
Step 3	Data Collection: List of bundles was created and analyzed in different steps
Step 4	Functional analysis of the lexical bundles was done
Step 5	Data Analysis: Findings were discussed and analyzed.

2.1. Contrastive Interlanguage Analysis (CIA)

Granger (1996, p. 295) defined CIA as "a methodology involving comparison of learner data with native speaker data (L2 vs L1) or the comparison of different types of learner data (L2 vs L2)". The current study aimed to compare L2 and L1 writers with a focus on their development in target norms. Native and non-native corpora were compared to two native corpora to investigate lexical bundle usage patterns. Regarding the significance of contrastive interlanguage analysis, Huang (2014) stated that it presents crucial information regarding the variations in learner English. Lado (1957, p. 1) indicated that "in the comparison between native and foreign language lies the key to ease or difficulty in foreign language teaching". CIA model has the potential to show us the performance problems in written productions of EFL learners. Kohn (1986, p. 21) observed that "transfer is one of the major factors which shapes learners' interlanguage performance and competence." It can be seen, therefore, that problems related to L1 transfer into L2 have impacts on L2 language performance and this situation may be one of the good reasons behind problematic word usages in the target language.

In addition to the contrastive interlanguage analysis, the analysis of bundle patterns in functional terms in non-native corpora was also done in the framework of the study. First of all, a taxonomy by Biber, Conrad and Cortes (2003) was used to classify lexical bundles and this taxonomy was used for determining the functional categories of the bundles. The categories of functional lexical bundles were first proposed by Cortes (2002). Biber et al. (2004) developed taxonomy later. This taxonomy contained stance expressions, discourse organizers and referential expressions. Stance expressions are described as "overt expression of an author's or speaker's attitudes, feelings, judgments" (Biber et al., 2004, p. 386). The second category of bundles is called "discourse organizers", which structure the texts with such sub-functions as "presentation, clarification and elaboration". The third category is "referential expressions", and they are used for a particular attribute or a condition. These categories are given below with examples of functional categories.

Table 3.
Lexical Bundles' Functional Categorization (Biber et al., 2004, p.384)

1. Stance Expressions	2. Discourse Organizers	3. Referential Expressions
A. Epistemic Stance	A. Topic Introduction/Focus	A. Identification/ Focus:
Personal: they do not	*of this study is*	*one of the most important*
Impersonal: *it can be a*		B. Imprecision:
		and things like that
B. Attitudinal/ Modality Stance	B. Topic	C. Specification of Attributes
B1) Desire: *if you want to*	Elaboration /Clarification *on the*	C1) Quantity Specification: *Variety of the*
B2) Obligation / Directive	*other hand*	C2) Tangible Framing Attribute:
Personal: *we want to learn*		*in the shape of*
Impersonal: *it is necessary to*		C3) Intangible Framing Attribute:
B3) Intention/Prediction		*in the case of*
Personal: *we can assume that*		
Impersonal: *is going to be*		D. Time/Place/Text Reference
B4) Ability		D1) Place Reference: *in the classroom*
Personal: *be able to*		D2) Time Reference: *in the same time*
Impersonal: *it can be*		D3) Text Deixis: *as it can be seen in Figure*
		D4) Multi-functional Reference: *In the beginning of*

Being a corpus query software, Sketch Engine is a popular free online corpus tool which is used for exploring the ways in which language behaves under specific conditions. The tool has the capacity to make analysis over a very large number of authentic data, showing what is typical, rare or unusual in linguistic terms. Linguists, translators, language teachers and learners and lexicographers frequently use it. Several investigations in the study were carried out using Sketch Engine software. Within the scope of the study, lexical bundles on four corpora were analyzed for frequencies and created the most common word lists of the target lexical bundles.

2.2. Expository Argumentation and Academic Writing (Corpora)

The distinction between the two types of writing is clear-cut. Expository argumentative writing is an essential part and form of essay writing that is based on the ability to form convincing arguments aiming to persuade the audience with strong and rational arguments, not necessarily dealing with academic topics. Academic writing is a form of writing that consists of a focused and structured form with a formal tone and style along with some general features which are relevant across all disciplines. The corpora used for this study consisted of several argumentative and academic corpora of essays that are produced by L1 and L2 writers. The written productions were compared on the basis of the frequency and variety of lexical bundle content based on functional aspects.

3. FINDINGS

The data collection and analysis processes were done in terms of several dimensions. First of all, the raw frequencies were obtained and then they were normalized. Their log-likelihood values and functional categories were created and presented in the form of tables and graphics below.

3.1. Common Bundles in BAWE and KTUCALE

Through a corpus-based contrastive analysis, the bundles were analyzed, and the commonest bundle patterns in BAWE and KTUCALE are given in Table 4 below. It is not surprising that native corpus provided important comparison data between the native and non-native bundles. Table 4 shows that L1 speaker use of bundle patterns seem more homogenous than those of L2 writers. With no prior exposure to language varieties, it seems that L2 learners seemed to have used fewer bundle patterns more frequently than their native counterparts. Furthermore, normalized frequency values of both groups were relatively high, indicating that the groups tend to use the same or similar bundles.

Table 4.
Frequent Bundle Patterns in BAWE and KTUCALE

BAWE			**KTUCALE**			
First 40 bundles in BAWE	**Raw frequency**	**Normalized frequency**	**Raw frequency**	**Normalized frequency**	**-,+**	**LL score**
the use of	374	599,08	239	291,52	-	78,11
in order to	325	520,59	527	642,80	+	9,07
be able to	231	370,02	300	365,92	-	0,02
there be a	186	297,94	300	365,92	+	4,92
the fact that	180	288,33	118	143,82	-	35,40
that there be	153	245,08	185	225,65	-	0,57

there be no	140	224,25	111	135,39	-	15,91
that it be	137	219,45	125	152,47	-	8,67
of the poem	135	216,25	3	3,66	-	200,93
one of the	134	214,64	556	678,18	+	175,00
way in which	133	213,04	39	47,57	-	83,09
the end of	129	206,63	64	78,06	-	43,61
part of the	129	206,63	140	170,36	-	2,42
it be not	124	198,63	208	253,71	+	12,77
as well as	123	197,02	176	214,67	+	0,56
use of the	121	193,82	45	54,90	-	59,91
it be a	121	193,82	135	165,64	-	1,69
in term of	121	193,82	237	289,65	+	13,29
a sense of	118	189,01	23	28,05	-	98,53
the way in	113	181,01	29	35,37	-	78,61
the way in which	110	176,20	26	31,71	-	81,23
can be seen	105	168,19	35	42,69	-	58,25
the importance of	104	166,59	178	217,11	+	4,70
due to the	103	164,99	44	53,67	-	43,16
on the other	102	163,39	261	318,35	+	35,45
seem to be	100	160,18	53	64,65	-	30,31
the other hand	99	158,58	259	315,91	+	37,13
on the other hand	99	158,58	255	311,03	+	35,21
it be the	96	153,77	94	114,64	-	4,08
be used to	96	153,77	21	25,61	-	74,67
be seen as	95	152,17	52	63,43	-	27,70
a number of	95	152,17	83	101,24	-	7,37
the idea of	93	148,97	49	59,77	-	28,48
cannot be	93	148,97	130	158,57	+	0,21
at the end	90	144,16	41	50,01	-	34,56
the role of	89	142,56	115	140,27	-	0,01
as it be	89	142,56	58	70,74	-	17,75
refer to the	86	137,76	41	50,01	-	30,91
look at the	86	137,76	35	42,69	-	38,72
the meaning of	84	134,55	190	231,75	+	18,28

The raw and normalized frequencies of the bundles show that there are bundles in each corpus such as one of the and in order to in KTUCALE and the use of and in order to in BAWE which are used more frequently, indicating that although preferences of L1 and L2 learners may differ, the most frequent bundles may be the same or similar ones. A note of caution is that as Schmitt and Carter (2004, p. 13) stated, due to the "lack of rich input", EFL learners may present overuse and underuse usage patterns which are also common in L2 writing. This indication is further strengthened by Li and Schmitt (2009), who found that L2 learners are likely to overuse the bundle patterns they are exposed to.

Another significant point is that although the most frequent bundles were generally the same in both corpora, there were also several diverging usage patterns. Such bundles as one of the, on the other, on the other hand, the other hand seemed to have been used to a considerable extent in KTUCALE although they are not often seen in BAWE. Yet, it is also seen that some bundles such as of the poem, way in which, the end of, seem to be which are used in BAWE were not used in KTUCALE. Since the BAWE corpus, and constitutes a 'norm', it may be right to speculate that non-native learners do not use the highest frequency bundles especially specific to the native speakers.

The analysis made for the distribution of the bundle patterns reveals that several bundle patterns are frequently used by L2 learners in surprisingly higher frequencies. This, in turn, indicates that L2 learners tend to overuse some bundles due to the limited exposure (Reppen & Biber, 2016). It can also be seen that both corpora show some contradictory results, indicating problems in the number and variety of non-native usage patterns.

There were only nine bundles used by L1 speakers like the L2 use, and this accounted for 18% of the bundles in total. 82% of them, however, were incompatible with the native usage patterns. Among them, the role of with a LL score of 0,01 is the most consistent one in both corpora. Such bundles as it has to be, as a result, cannot be, it be a, as well as, part of the, be able to and that there be were the other samples that were consistent with the reference corpus. These bundle patterns shared the common features of consisting of three words.

It is also important to note that there were lots of underuse and overuse patterns used by L2 learners. One good explanation for this may be the use lexical bundles which are less diverse and more limited in the essays compared to those of native learners since the amount exposure is greater and longer. According to Adel and Erman (2012), L2 learners tend to use bundle

patterns which were fewer and limited in comparison to native speakers. Nesselhauf (2005) emphasized liability of L2 learners to underuse patterns, and Ellis (2012) added that L2 learners tend to use the bundle patterns that are very familiar to them. The data also shows that the frequently used bundles by L1 speakers are mostly underused by the L2 learners. The highest frequency of underuse patterns seemed to be use of the, way in which, a sense of, the way in which, be used to and an example of. In terms of overuse bundles, Salazar (2006, p. 134) noted that "further examination of the overused bundles indicates the non-native writers' excessive reliance on a handful of highly frequent bundles, to the detriment of less common bundles with similar meanings". Accordingly, there were twelve (12) bundle patterns in Table 4 that had overuse patterns by non-native learners. One example for this was one of the which was the most frequently overused item with the LL score of 175. The second highest overused item was the other hand with the LL score of 37,13. It is also clear that the bundle patterns which are frequently used are used at least three times more. Besides, in 12 overused bundles, 6 of them were considerably overused by the L2 learners; these being it can be, the meaning of, the other hand, one of the, be seen as and on the other hand. The possible reasons behind the overuse and underuse patterns in non-native data may be due to such reasons as lack of knowledge, self-confidence, lack of training and the effect of first language. Paquot (2013, p. 402) argues that "the first language may prompt learners to use lexical bundles that display untypical patterns in English". Another reason for the overuse of data may be the language transfers from L1 and to support this, Granger (2014. p. 69) also argued that "the lack of salience that characterizes many lexical bundles constitutes a challenge for learners and trainees who may be led to produce awkward-sounding phrases, often directly transferred from their mother tongue".

3.2. Common Bundles in TICLE and LOCNESS

The biggest number of frequently used bundle patterns in LOCNESS and TICLE, listed according to their normalized frequencies and LL scores are given in Table 5. When the list of three-word lexical bundles in both corpora is extracted, it is seen that the three-word lexical bundles with the highest frequencies in TICLE were do not have (702), they do not (1083), a lot of (555), in order to (528), one of the (470), cannot be (434) and should not be (366). Some lexical bundles were also found as frequent lexical bundles in previous studies. 23 out of 40 bundles turned out to be more than one hundred times in one-million word and 12 bundles were found to be belonging to the list of the most frequent bundle patterns by Biber et al. (2004), and Cortes (2008). LOCNESS corpus, however, displayed rather different usage patterns in Table 5. The most frequent three-word lexical bundles in LOCNESS were able to (520), the fact that (448), in order to (351), one of the (335) and they do not (252). LOCNESS presented rather different results, but many bundles it contained were the same as the previously identified bundles by Biber et al. (2004) and Cortes (2008). There were overuse and underuse patterns in the two non-native corpora (TICLE and KTUCALE) when compared to the native ones. The total amount of underused bundles reached 34 %, and it seemed that there was a relatively balanced distribution between the two corpora. With this in mind, however, the rate of underuse patterns in TICLE and LOCNESS was higher than those in academic corpora (BAWE and KTUCALE). The reasons that affect the use of L1 bundle usage patterns may be given to several factors such as that L1 speakers may be exposed to underused bundles in lectures (Krishnamurthy, 2002),

In Table 5, it is seen that there were also overuse pattern problems that we came across mostly in L2 speakers' writing (Hyland, 2008a; Kamshilova, 2017; Pan et al., 2016). Such bundles as on the other, a lot of, do not have and they do not were overused and they do not was on the top of the list with the LL score of 162,48 and frequency of 1083,02 per million. However, LOCNESS data shows that its normalized frequency is just 252,03 times per million.

Table 5.
Frequent Bundle Patterns in TICLE and LOCNESS

Bundles	LOCNESS		TICLE		LL score	
	Raw frequency	Normalized frequency	Raw frequency	Normalized frequency	- ,+	
be able to	188	520,70	75	335,65	-	10,94
the fact that	162	448,69	32	143,21	-	43,88
in order to	127	351,75	118	528,08	+	9,98
that it be	124	343,44	74	331,17	-	0,06
one of the	121	335,13	105	469,91	+	6,34
they do not	91	252,03	242	1.083,02	+	162,48
the end of	82	227,11	24	107,41	-	11,76
the idea of	81	224,34	19	85,03	-	18,98
have to be	80	221,57	28	125,31	-	7,71
because of the	79	218,80	58	259,57	+	0,97
this be a	76	210,89	34	152,16	-	2,57
due to the	76	210,89	7	31,33	-	38,67
that they be	75	207,73	56	250,62	+	1,12
the right to	73	202,19	38	170,06	-	0,67
end of the	73	202,19	12	53,70	-	24,71
the death penalty	71	196,65	6	26,85	-	37,80
should not be	71	196,65	82	366,92	+	14,79

the use of	70	193,88	22	98,46	-	8,54
the number of	70	193,88	30	124,36	-	2,96
of the world	69	191,11	51	228,24	+	0,92
cannot be	68	188,34	97	434,10	+	28,45
the end of the	67	185,57	11	49,23	-	22,24
part of the	67	185,57	47	210,34	+	0,43
in the united	67	185,57	4	17,90	-	41,46
be not the	67	185,57	34	152,16	-	0,*91*
do not have	65	180,03	157	702,62	+	96,10
in the world	64	177,26	108	483,33	+	42,30
for the good	64	177,26	2	8,95	-	47,68
to be a	63	174,49	39	174,54	+	0,00
as well as	63	174,49	20	89,51	-	7,50
a lot of	61	168,95	124	554,94	+	62,67
be one of	60	166,18	65	290,89	+	9,73
be in the	60	166,18	44	196,91	+	0,62
that he be	59	163,41	9	40,28	-	21,00

The total amount of underused bundles equals to 34 % and it seems that there is relatively a balanced distribution in both corpora (LOCNESS and TICLE). However, the rate of underuse patterns was higher than those in academic corpora (BAWE and KTUCALE).

3.3. Shared Lexical Bundles in Six Corpora

In the previous section, the analysis of bundle patterns in native and non-native corpora were made to determine the usage patterns in terms of academic and expository argumentative writing. In this section, the analysis is extended to include the other corpora by LSWE and AUGER. The rationale for doing so was to measure the extent of possible convergence between the non-native and native corpora. The first ten most common bundle patterns were obtained from the three reference corpora, and a frequency comparison was made between native and non-native corpora in order to see the possible convergent or divergent patterns.

Table 6.
Ten Most Frequent Bundle Patterns in Biber's and Davis' Corpora as well as in KTUCALE and BAWE

	KTUCALE Norm.	**BAWE Norm.**	**LSWE Norm.**	**AUGER Norm.**	**TICLE Norm.**	**LOCNESS Norm.**
one of the	685,50	217,85	200+ above	277,82	474,38	340,67
on the other hand	376,19	158,58	100+above	121,73	469,42	138,46
on the other	318,35	163,39	100+above	146,08	532,56	160,64
in terms of	282,98	193,82	100+above	186,08	80,56	52,62
as well as	259,80	197,27	200+above	293,04	89,51	210,49
one of the most	224,02	28,83	100 -below	62,60	196,94	85,86
part of the	170,76	206,63	200+above	185,65	210,34	185,57
the fact that	150,03	291,53	200+above	151,30	143,21	451,46
of the most imp.	136,67	4,81	100-below	26,95	8,95	27,70
the effect of	120,75	102,52	100-below	92,17	62,65	94,17

It can be seen in Table 6 that there are overuse patterns by L2 learners with one of them and it is significant that some lexical bundle patterns are overused. Bundles such as one of the, on the other hand, on the other, in terms of, one of the most, and of the most important are the overuse patterns while the bundles such as, the fact that and part of the are underused. The normalized frequencies of the several bundles in L2 corpora seem to be more than those in L1 corpus in spite of the fact that some bundles were used more by L1 writers. For instance, KTUCALE showed an interestingly high percentage for the use of the most important bundle with a normalized frequency of 136.67. This frequency is higher than all frequency values in the table for the same bundle. Possible reasons for this overuse pattern are given in the next section. TICLE, however, showed a balanced distribution with the other corpora and even slightly underused patterns were observed when compared to the AUGER and LOCNESS.

Another purpose of the analysis was to categorize the lexical bundles' functional features. For this reason, the analysis of KTUCALE to determine the functional features was made with a focus on the taxonomies introduced earlier. For space considerations, the functional analysis of LOCNESS was not included in this current study.

3.4. Lexical Bundles in KTUCALE and Functional Analysis

The bundle patterns found in KTUCALE were functionally analyzed, and the related categorizations were made in terms of several categories. These were discourse organizers, referential expressions and stance expressions with their subcategories which can be observed in Table 7. The taxonomy of functional categorization by Biber et al. (2004) was a perfect fit for the lexical bundles found in KTUCALE. Each category is exemplified by the related bundles from the corpus under study.

First of all, stance expressions used to express personal feelings, attitudes, perspectives, certainties, and uncertainties (Biber, 2006) were grouped under two major categories: these being epistemic stance and attitudinal modality stance bundles. In general terms, the percentage of stance bundles was limited to a few repeated bundles (22%) in the most frequent 100 bundles in KTUCALE. Compared to other groups, the stance category consisted of such epistemic stance bundles as they do not, they be not, they cannot, we cannot, you do not, I think that, do seem to be, is likely to be, the fact that the. Some of the attitudinal stance bundles that show the personal viewpoints included if you have, you want to, if we want to, I want you to, I am not going to, it should be, have to be, it is necessary to and I am going to.

Table 7.
Functional Categorization of Lexical Bundles' in KTUCALE (Biber et al., 2004, p. 384)

1. Stance Expressions
A. Epistemic Stance
Personal: *they do not, they be not, they cannot, we cannot, you do not, I think that*
Impersonal: do *seem to be, is likely to be, the fact that the*
B. Attitudinal/ Modality Stance
B.1) Desire: if you have, you want to, if we want to
B.2) Obligation/ Directive
Personal: *I want you to, I am not going to*
Impersonal: *it should be, have to be, it is necessary to*
B.3) Intention/Prediction
Personal: *I am going to*
Impersonal: *are going to be,*
B.4) Ability
Personal: *I am able to, that we can*
Impersonal: *to be able to*
2. Discourse Organizers
A. Topic Introduction/Focus: *in this chapter we, this part of the study*
B. Topic Elaboration/Clarification: *on the other hand, for the purpose of, on the part of, in the same way*
3. Referential Expressions
A. Identification/ Focus: *one of the most, one of the, the most important, be one of,*
B. Imprecision: *and things like that*
C. Specification of Attributes
C.1) Quantity Specification: *a variety of, a lot of, there be many, there is no,*
C.2) Tangible Framing Attribute: *in the form of, the results of the, this study is to, is found to be*
C.3) Intangible Framing Attribute: *on the basis of, the aim of this, the nature of the, with the help of*
D. Time/Place/Text Reference
D.1) Place Reference: *in the United States*
D.2) Time Reference: *at the same time, the end of the*
D.3) Text Deixis: *in the present study, at the beginning of,*
D.4) Multi-functional Reference *is related to the, the first and second*

The discourse organizers category helps to introduce, discuss and clarify a point included a few lexical (16%) bundles such as in this chapter we, this part of the study, on the other hand, for the purpose of, on the part of, in the same way, etc.

Referring to physical or abstract entities (Biber et al., 2004), the final group of lexical bundles is referential expressions which include four main sub-categories; these being identification, imprecision, specification, and time/place/text references. In general terms, the percentage of the total number of referential expressions in the most frequent 100 bundles in KTUCALE were found to be 62%. Compared to other groups, the referential expression category consisted of many examples from the KTUCALE. The examples were given based on the four subcategories. The first is the "identification" category with such examples as one of the most, one of the, the most important, and be one of. The second subcategory is "imprecision" such as and things like that. The third subcategory was "specification" and it included such examples as a variety of, a lot of, there be many, there is no, on the basis of, the aim of this, the nature of the, and with the help of. The final category was "time/place/text references" with examples such as in the United States, at the same time, the end of the, in the present study, at the beginning of, is related to the, and the first and second.

This section has attempted to make a functional bundle analysis in KTUCALE revealing that the "referential expressions" category was found to be the largest lexical bundle category, with a percentage of 62%. "Stance bundles" and "discourse organizers" categories were used with a percentage of 38% in the corpus.

4. RESULTS, DISCUSSION AND RECOMMENDATIONS

In this corpus-based contrastive analysis, the purpose was to analyze the phraseological awareness of native and non-native writers. The focus was particularly on comparing the bundle usage patterns in native and non-native learners, and the results showed that bundles were employed homogeneously in academic and expository argumentative texts by both groups of non-native writers.

The first question was about the characteristic elements of lexical bundles in the essays, and it was seen that Turkish EFL learners avoided using common bundles, and their usage patterns were limited to a few but repeated lexical bundles. The probable rationale for this situation may be their use of the "avoidance" strategy. For Granger (1998), De Cock (2000), and Foster (2001), the repetitive use of lexical bundles by EFL learners is due to the fact that they are considered as the most reliable recurrent expressions to be used in any context. When learners use the same or very similar bundle patterns, this may increase their confidence in writing. However, it is also the case that this practice may also cause overuse problems. According to Laufer (2000), the strategy of using repetitive elements by EFL learners may be for overcoming problems that appeared due to incongruences in between the usage patterns of L1 and L2 and the lack of exposure to native bundle usage patterns. Cortes (2004) also supports these findings, claiming that non-native writers favored repetition of the same bundles in their written productions.

The second question was about whether there were any differences of lexical bundle performance in both groups. It became clear that l2 writers largely employed few bundle patterns, and they did not use many frequent bundles found in the native corpus. The possible reasons for this may be given to several factors such as insufficient exposure to a variety of lexical bundles in a L2 context, and the texts they are exposed to are produced by experts in L2 context. It is obvious that non-native learners should be encouraged to use authentic materials. According to Cortes (2004), noticing activities can be used to improve awareness of functions, structures and context of lexical bundles and this should also be recommended for each register. It was also the case that bundles used by Turkish EFL learners are not those frequently used by native writers.

The third research question was about the bundle patterns that are much preferred by Turkish EFL learners. The preferred bundles in the non-native corpus were to some extent like those in the native corpus. However, it was also the case that the normalized frequencies in the non-native corpus revealed that several bundle patterns were used more than those in the native speaker corpus.

The fourth question was about performance differences of bundle patterns in the written productions of Turkish EFL learners. In KTUCALE and TICLE, bundle usage patterns varied. There was almost no difference for the most common bundle preferences, but differences were noted when the other bundle patterns were analyzed. There were several common and frequent bundles in the native corpus that were not preferred by the learners in both corpora, and their log-likelihood scores and normalized frequencies differed a lot. In comparison with the native corpus, KTUCALE presented some underuse patterns, and the distribution was balanced in terms of overuse patterns in TICLE when they are compared to the equivalent native corpus. The analysis of the common bundle patterns is done and compared with the reference corpus and it was seen that TICLE corpus contained more common bundles than KTUCALE.

The final question was about the functional categories of lexical bundles in KTUCALE. "Referential expressions" was found to be the highest category of the functional categories in KTUCALE. This means that the essays were written in descriptive character. Past research brought about different results regarding the use of bundle patterns according to a taxonomy. Chen and Baker (2016) found that the "discourse organizers" category was the most functional category in BAWE-CH. However, Ädel and Erman (2012) found functional referential expressions in their study. Another finding was that the biggest part of the proportion was functioning as referential expressions in the essays from the Stockholm University Student English Corpus (SUSEC). The researchers found similar proportions of referential expressions from the learner corpus (SUSEC) and the native speaker corpus. In her study of "Lexical Bundles in Academic Texts by Non-native Speakers", Dontcheva-Navratilova (2012) found that the distribution of referential expressions was slightly higher than in other categories but differed considerably from the conventions of expert academic writing. She noticed that novice writers were using a restricted number of lexical bundles in academic writing, partly because of the insufficient level of rhetorical skill development. Past research also suggests that the most functional types of bundle patterns in academic prose were the referential bundles (Biber and Barbieri, 2007). In social sciences, the category of discourse organizer is one of the most noticeable functions of the lexical bundle. This is because social sciences are “the more discursive and evaluative patterns of argument in the soft knowledge fields, where persuasion is more explicitly interpretative and less empiricist” (Hyland 2008a, p. 16). The results of many similar studies related to the lexical bundles suggest that referential bundles are the preferred bundles, which indicates that “novice writers have not yet acquired discipline-specific discourse conventions” (Dontcheva-Navratilova, 2012, p. 55).

This corpus-based contrastive study was conducted to investigate tertiary level Turkish EFL learners' lexical bundle usage patterns. The analysis was made through functionally analyzing and identifying the frequencies, overuse and underuse patterns of the bundle patterns from native and non-native learner corpora on L2 English expository argumentative and academic essays. Both non-native corpora were designed according to strict design criteria and contained words slightly over one million, having almost 820.000 words for the academic corpus. The novelty of this study is highlighted in three steps. First, the findings suggest that there is a need to integrate lexical bundles into the curriculum and coursebooks that are specially designed for Turkish EFL learners. The integration of the most frequent lexical bundles as well as their usage patterns and contextual relationships should necessarily be highlighted for novice learners who can easily pick up these word patterns, which can also help determine the extent of pedagogical intervention required. Secondly, it is possible that the bundles can be ordered by frequency, form, and function. If Turkish EFL learners are given more opportunities and informed feedback related to the lexical bundles which are carefully selected from the naturally occurring data through the native corpora, their awareness may be increased, thus learning to use appropriate lexical bundle patterns to serve the right function in each context. Thirdly, as part of the efforts to improve materials design and curriculum development, it is possible to integrate bundle patterns into the content of reading and listening activities in the coursebooks, and explicit teaching activities based on the form and function of lexical bundles can be planned for Turkish EFL learners. Particularly, it seems that there is a need for integrating the most common lexical bundle patterns into the writing curriculum for fostering proper use. In other words, explicit teaching of the most common lexical bundles should be integrated into the writing curriculum, and writing instructors and material writers should emphasize these word groups in their teaching and coursebooks.

This study contributes to the existing literature in that it helps increase the understanding of the combinational nature of the English language by focusing on the bundle patterns in the argumentative essays of tertiary level EFL writers, contributing to a better understanding of idiomatic principles of English lexicon by the Turkish-English writers. The study can also inspire teachers and material writers to incorporate the most frequently and commonly employed lexical bundles into their teaching and coursebook content. One limitation was that the corpora studied were very big, and lots of instances were left out of the scope of the study for space considerations.

Research and Publication Ethics Statement

This study has not been presented in any congress or symposium. In addition, it has not been sent to any other journal for publication. This study is partly based on the first author's MA Thesis in 2019.

Contribution Rates of Authors to the Article

The authors contributed equally to the study.

Statement of Interest

There is no conflict of interest from the authors to declare.

5. REFERENCES

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes*, 31, 84-91. https://doi.org/10.1016/j.esp.2011.08.004

Alsop, S., & Nesi, H. (2009). Issues in the development of the British Academic Written English (BAWE) corpus. *Corpora,* 4(1), 71–83. https://doi.org/10.3366/E1749503209000227

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

Bal, B. (2010). *Analysis of four-word lexical bundles in published research articles written by Turkish scholars* [Unpublished master's thesis]. Georgia State University.

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405. https://doi.org/10.1093/applin/25.3.371

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. John Benjamin Publishing Company. https://doi.org/10.1075/scl.23

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275-311. http://dx.doi.org/10.1075/ijcl.14.3.08bib

Çelebi, D. (2006). Türkiye'de anadili eğitimi ve yabancı dil öğretimi. *Erciyes Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 21*(2), 285-307. https://dergipark.org.tr/tr/pub/erusosbilder/issue/23754/253119

Chang, C., & Kuo, C. (2011). A corpus-based approach to online materials development for writing articles. *English for Specific Purposes*, *30*(3), 222-234. https://doi.org/10.1016/j.esp.2011.04.001

Chen, Yu-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning and Technology, 14*(2), 30-49. http://dx.doi.org/10125/44213

Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes, 23*(4), 397-423. https://doi.org/10.1016/j.esp.2003.12.001

Cowie, A. P. (Ed.). (1998). *Phraseology: Theory, analysis, and applications*. Oxford University Press.

Djigunović, J. M., & Krajnović, M. M. (Eds.). (2012). *UZRT 2012: Empirical studies in English applied linguistics*. FF Press.

Dontcheva-Navratilova, O. (2012). Lexical bundles in academic texts by non-native speakers. *Brno Studies in English, 38*(2), 37-58. https://doi.org/10.5817/BSE2012-2-3

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics, 32*, 17-44. https://doi.org/10.1017/S0267190512000025

Erman, B. (2009). Formulaic language from a learner perspective: What the learner needs to know. In B. Corrigan, H. Quali, E. Moravcsik, & K. Wheatley (Eds.), *Formulaic language* (pp. 27–50). John Benjamins.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (Ed.), *Phraseology: theory, analysis, and applications* (pp. 145–160). Oxford University Press.

Granger, S. (1998). The computerized learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3- 18). Longman.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics in Language Teaching, 52*(3), 229-252. https://doi.org/10.1515/iral-2014-0011

Granger, S. (2014). A lexical bundle approach to comparing languages: Stems in English and French. *Languages in Contrast, 14*(1), 58-72. https://doi.org/10.1075/lic.14.1.04gra

Güngör, F., & Uysal, H. H. (2016). A comparative analysis of lexical bundles used by native and non-native scholars. *English Language Teaching, 9*(6), 176-188. https://doi.org/10.5539/elt.v9n6p176

Hoey, M. P. (2000). *Patterns of lexis in text*. Oxford University Press.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44. https://doi.org/10.1093/applin/19.1.24

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524773

Hyland, K. (2008a). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, K. (2008b). *Academic clusters: Text patterning in published and postgraduate writing. International Journal of Applied Linguistics, 18*(1), 41-62. https://doi.org/10.1111/j.1473-4192.2008.00178.x

Kamshilova O.N. (2017). Overuse in learner language: Frequency and accuracy. *Russian Linguistic Bulletin, 3*(11), 28-31. https://dx.doi.org/10.18454/RULB.11.12

Kilimci, A., & Can, C. (2009). TICLE: Uluslararası Türk öğrenici İngilizcesi derlemi. In M. Sarıca, N. Sarıca & A. Karaca (Ed.), *XXII. Ulusal Dilbilim Kurultayı Bildirileri* (pp. 1- 11). Ankara: Yüzüncü Yıl Üniversitesi Yayınları.

Krishnamurthy, R. (2002). Language as chunks, not words. JALT2002: *Conference proceedings*, 288-294. https://jalt-publications.org/archive/proceedings/2002/288.pdf

Lado, R. (1957). *Linguistics across cultures: Applied linguistics for language teachers.* University of Michigan Press.

Laufer, B. (2000). Avoidance of idioms in a second language: The effect of L1-L2 degree of similarity. *Studia Linguistica, 54*(2), 186-196. https://doi.org/10.1111/1467-9582.00059

Lehmann, M. (2013). The use of lexical bundles in EFL academic writing tasks. In J. M. Djigunović & M. Krajnovic (Eds.), *UZRT 2012: Empirical studies in Magdolna Lehmann* (pp.131-141).

Meunier, F., & Granger, S. (2008). *Phraseology in language learning and teaching.* John Benjamins. https://doi.org/10.1075/z.138

Miller, N. (2009) Assessing the processing demands of learner collocation errors. *Poster presented at Corpus Linguistics Conference 2009*, Liverpool, U.K.

Nesselhauf, N. (2005). *Collocations in a learner corpus.* John Benjamins. https://doi.org/10.1075/scl.14

Pan, F., Reppen, R., & Biber, D. (2016). Comparing patterns of L1 versus L2 English academic professionals: Lexical bundles in Telecommunications research journals. *Journal of English for Academic Purposes, 21*, 60-71. https://doi.org/10.1016/j.jeap.2015.11.003

Paquot, M. & Granger, S. (2012), Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32*, 130-149. https://doi.org/10.1017/S0267190512000098

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: native like selection and native like fluency. In J. Richards & R. Schmidt (Eds.), *Language and communication* (pp. 191-226). Longman.

Romer, U. (2010). Establishing the phraseological profile of a text type: the construction of meaning in academic book reviews. *English Text Construction, 3*(1), 95-119. https://doi.org/10.1075/etc.3.1.06rom

Ruan, Z. (2017). Lexical bundles in Chinese undergraduate academic writing. *RELC Journal, 48*(3) 327–340. https://doi.org/10.1177/00336882166312

Salazar, D. J. L. (2011). *Lexical bundles in scientific English: A corpus-based study of native and non-native writing* [Unpublished Ph.D. dissertation]. Universitat de Barcelona.

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford University Press.

Tracy, S. D. (2012). *Undergraduate vs. graduate academic English: A corpus-based analysis* [Unpublished doctoral dissertation]. The Pennsylvania State University.

Öztürk, Y., & Köse, G. D. (2016). Turkish and native English academic writers' use of lexical bundles. *Journal of Language and Linguistic Studies, 12*(1), 149-165. https://files.eric.ed.gov/fulltext/EJ1105228.pdf

Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal, 42*(2), 155-166. https://doi.org/10.1177/00336882114072

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language & Communication, 20*(1), 1–28. https://doi.org/10.1016/S0271-5309(99)00015-4