ISSN 2335-2019 (Print), ISSN 2335-2027 (Online)
Darnioji daugiakalbystė | Sustainable Multilingualism | *26*/2025
https://doi.org/10.2478/sm-2025-0005

Sibel Aybek
Cukurova University, Türkiye
Cem Can
Cukurova University, Türkiye

COMPARATIVE ANALYSIS OF LEXICAL BUNDLES IN ACADEMIC WRITINGS BY NATIVE ENGLISH SPEAKERS AND TURKISH EFL LEARNERS

Abstract. Authentic language use frequently consists of repeated expressions called multiword units or formulaic utterances (Byrd & Coxhead, 2010), which serve as essential "building blocks of discourse in both spoken and written registers" (Biber & Barbieri, 2007, p. 263). Lexical bundles, a subset of formulaic sequences, are defined as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber et al., 1999, p. 990). This study investigates the use of the most frequent 3- and 4-word lexical bundles in the TICLE, the Turkish component of the International Corpus of Learner English (ICLE), and the Louvain Corpus of Native English Essays (LOCNESS) as the control parallel corpus. The lexical bundles are classified according to their structural and functional characteristics based on the taxonomy developed by Biber et al. (2003; 2004). An interpretative contrastive analysis was conducted between the native (LOCNESS) and non-native (TICLE) data sets. The findings reveal that Turkish EFL learners overuse verb phrase fragments while underusing noun phrase and prepositional phrase fragments. Furthermore, texts in TICLE exhibit a lower lexical variety compared to those in LOCNESS. Regarding functional classification, although Turkish EFL learners produce fewer functional bundles overall, they tend to overuse a limited subset of them. These results suggest underlying issues in EFL pedagogy, particularly the need for explicit instruction on multiword units.

Keywords: corpus linguistics; learner corpora; lexical bundles; multiword units; Turkish EFL learners.

Introduction

Multi-word units "are important building blocks of discourse in spoken and written registers" (Biber & Barbieri, 2007, p. 263). These lexical sequences are integral to both oral and written language production and processing,

providing "a steppingstone into language development" (Chenu & Jisa, 2009, p. 27). It is emphasized that the studies on the use of multi-word units in the written performances of English language learners have significantly influenced lexicography and English language textbooks. These studies shed light on the authentic uses of multi-word units, providing insights into their structures and functions within discourse, and inform learners, teachers, and material developers alike.

From a psycholinguistic perspective, formulaic utterances offer "a processing advantage over creatively generated language" for non-native speakers (NNS) and native speakers (NS) (Conklin & Schmitt, 2008, p. 72). Similarly, Jiang and Nekrasova (2009) present "prevailing evidence in support of the holistic nature of formula representation and processing in second language speakers" (p. 433).

Learner corpus studies, which involve the systematic analysis of language produced by learners, have provided invaluable insights into second language acquisition and usage patterns. Lexical bundles, a specific focus within this field, are recurrent sequences of words that serve key functions in discourse, despite not always being idiomatic or syntactically complete. Pioneering research by Biber et al. (2004) and Granger (1998) has highlighted the importance of these sequences in understanding language proficiency.

The current research provides many insights for both SLA researchers, English language teachers, instructors, material developers and publishers as it explores how Turkish EFL learners produce multiword units in their written performances, providing English language teachers with insights into the nature of these units. These structures not only reflect EFL learners' English proficiency levels but also offer teachers valuable input for creating effective materials and activities to enhance Turkish EFL learners' lexical competences.

Present study offers interpretations on how Turkish EFL learners produce multiword combinations in their written performances and helps English language teachers understand the nature of these lexical bundles. Acquisition and production of these items also reveals the necessity of using corpora and authentic data both in and out of the classroom as recommended by Sinclair (1998) underlining that, "[c]orpora will clarify, give priorities, reduce exceptions and liberate the creative spirit" (p. 38) for the learner.

The primary aim of this study is to investigate the use of 3- and 4-word lexical bundles in the TICLE learner corpus, which contains argumentative essays written by Turkish EFL learners. The study focuses on the most frequently used expressions, examining their structures and functions in the written performances of Turkish university students. The goal is to deepen our understanding of the use of these units. Following the identification of the structural and functional features of the lexical bundles used by Turkish EFL learners, similarities and differences are analyzed through a comparison with the use of these sequences by native speakers in LOCNESS, a reference corpus containing written essays by American university students (Granger et al., 2009).

This study addresses the following research questions:

1. What are the most frequent 3- and 4-word lexical bundles found in argumentative essays of Turkish EFL learners in the TICLE and LOCNESS corpora?
2. How are these lexical sequences in the TICLE and LOCNESS corpora classified based on their structural and functional characteristics?
3. How do the most frequent 3- to 4- lexical sequences used by Turkish EFL learners compare to those found in the LOCNESS corpus in terms of structure, function, and diversity in their use?

This study makes a significant contribution to sustainable multilingualism by addressing the challenges and potential solutions for enhancing academic writing skills among non-native English speakers, specifically Turkish EFL learners. By examining the structural and functional characteristics of lexical bundles, the research uncovers patterns of overuse and underuse that reveal critical gaps in learners' language proficiency. These insights are directly relevant to developing pedagogical strategies that support the acquisition and development of plurilingual competence. The findings emphasize the importance of targeted, corpus-based instruction that not only addresses linguistic gaps but also fosters a more nuanced understanding of multilingual dynamics. Aiming to promote language-sensitive teaching, this study advocates for instructional approaches that embrace cross-linguistic influences and intercultural dialogue. The research aligns with the goals of

sustainable multilingualism by offering data-driven insights that can inform language policy, curriculum design, and didactic practices aimed at preserving linguistic diversity while equipping learners to navigate multilingual environments effectively.

Lexical Bundles

Biber et al. (1999) describe lexical bundles as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" and as "simply sequences of word forms that commonly go together in natural discourse" (p. 990). Examples of lexical bundles include expressions such *as a result of*, *take a look at*, *on the other hand*, *and I don't know*, *to be able to*, *are you going to*, *I don't know if*, *at the same time, have a lot of,* and *you know it was*.

Despite numerous studies on formulaic language over the past decade, Nekrasova (2009) suggests that further research is required to closely examine the structural and functional characteristics of lexical bundles. Most of studies on lexical bundles have compared their usage in expert and non-expert writing, with a particular focus academic writing.

Previous research has established that conversation and academic prose exhibit distinct patterns of lexical bundles (Biber & Barbieri, 2007; Biber, Conrad, & Cortes, 2003, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999). These studies revealed that while most bundles in conversation are clausal, many bundles in academic prose are phrasal. Biber, Conrad and Cortes (2004) argue that everyday language comprises multi-word prefabricated expressions (e.g., *if you see, in a nutshell, what I mean*) and that language is not "strictly compositional" (p. 372).

Biber, Conrad, & Cortes (2004) have categorized lexical bundles according to their structures and functions within discourse. This present study adopts the structural and functional taxonomy developed by Biber, Conrad, & Cortes (2004, p. 381–396) as summarized in Table 1.

The first structural type includes verb phrase fragments. These bundles begin with a subject pronoun followed by a verb phrase, as in *I'm not going to, that's one of the, it's going to be,* and *this is a*. Alternatively, these bundles

may begin directly with a VP, such as *take a look at, is going to be,* or question fragments like *how many of you* and *are you going to.*

The second major structural type resembles the first category in incorporating verb phrase elements, however, these bundles also contain dependent clause fragments. Examples include *I want you to, what I want to,* and *if we look at.* An additional sub-category has been propounded to the existing classification by the researchers, which is *other adverbial clause fragments* added to the dependent clause fragments category.

Table 1

Structural Classification of Lexical Bundles

Verb Phrase Fragments	a. 1st/2nd/3rd person pronoun + VP fragment	*I'm/it's going to*
	b. Discourse marker + VP fragments	*I mean I don't*
	c. Verb phrase with non-passive verb	*have a lot of*
	d. Verb phrase with passive verb	*is based on*
	e. Yes/no question fragments	*do you want to*
	f. WH-question fragments	*what do you think*
Dependent Clause Fragments	a. 1st/2nd/3rd person pronoun + dependent clause fragments	*I don't know if*
	b. WH-question fragments	*what I want to*
	c. *If*-clause fragments	*if you want to*
	d. *To*-clause fragments	*want to do this*
	e. *That*-clause fragments	*that I want to*
	f. Other adverbial clause fragments	**as I know**
Noun Phrase and Prepositional Phrase Fragments	a. Noun phrase with of-phrase fragment	*one of the things*
	b. Noun phrase with another post-modifier frr.	*those of you who*
	c. Other noun phrase expressions	*and stuff like that*
	d. Prepositional phrase expressions	*at the end of*
	e. Comparative expressions	*as well as the*
	f. Quantifier expressions	***more important than***

Note. Adapted from Biber et al., 2004, p. 381–396.

The last main structural type includes only phrasal components. Most of these bundles comprise noun phrase (NP) components preceding a post modifier, as in *the end of the, the way in which, those of you who* and *a little bit about.* The remaining bundles in this type consist of prepositional phrase (PP) components with embedded modifiers, such as *of the things that, as well as the, at the end of*. In addition, another sub-category has been added to the NP fragments, which is *quantifier expressions.*

In their study, Biber et al. (2004, p. 389–396) identified three primary discourse functions for lexical bundles in English: 1) stance expressions, 2) discourse organizers, and 3) referential expressions.

According to Biber et al. (1999), stance bundles form a frame for expressing attitudes or assessment of certainty and can be categorized into two*; epistemic* or *attitudinal / modality*. Discourse organizers coordinate the flow of ideas by establishing relationships between preceding and forthcoming discourse, assisting in the introduction, elaboration, and clarification of topics. Examples include *I want to talk about, if you look at* and *going to talk about*. These bundles serve two primary functions: topic introduction/focus and topic elaboration/clarification (Biber et al., 1999). Lastly**,** referential bundles encompass a wide range of lexical bundles that typically refer to textual context or physical or abstract entities, such as *the nature of the* and *that's one of the.* This category includes four sub-categories: *identification/focus, imprecision, specification of attributes,* and *time/place/text reference* (Biber et al., 2004).

Table 2

Functional Classification of Lexical Bundles

Stance bundles	A. Epistemic stance	*I don't know what*
	B. Attitudinal/Modality stance	
	a. Desire	*I don't want to*
	b. Obligation/Directive	*It is important to*
	c. Intention/Prediction	*I'm not going to*
	d. Ability	*to be able to*
Discourse organizers	A. Topic introduction	*if you look at*
	B. Topic elaboration/Clarification	*on the other hand*
Referential bundles	A. Identification/ Focus	*of the things that*
	B. Imprecision	*or something like that*
	C. Specification of attributes	
	a. Quantity specification	*there's a lot of*
	b. Intangible framing	*in the form of*
	c. Tangible framing	*in the case of*
	D. Time/ Place/ Text reference	
	a. Place reference	*in the United States*
	b. Time reference	*at the same time*
	c. Text-deixis	*as shown in figure*
	d. Multi-functional reference	*the beginning of the*

Note. Adapted from Biber et al., 2004, p. 381–396.

Previous Studies on Lexical Bundles

Studies on lexical bundles have focused on both written and spoken discourse (Biber et al., 1999; Biber, Conrad & Cortes, 2004). The T2K-SWAL Corpus, compiled from TOEFL 2000 corpus, consists of texts from academic

life, including classroom teaching, textbooks, study groups, and university catalogues, sampled from six major academic disciplines: business, engineering, humanities, social and natural sciences, and education (Biber et. al, 2004). Their study revealed that lexical bundles were used twice as often in classroom teaching as in conversation and four times as often as in textbooks. Lexical bundles are much more common in both conversation and classroom teaching than in the written registers.

Biber & Barbieri (2007) found that lexical bundles differ from other lexico-grammatical structures in their physical mode (spoken/written). Other studies focused on bundles in professional and novice writings. For example, Cortes (2004) examined the bundles in the written production of published authors and student essays in the fields of History and Biology. She found that students could not effectively use lexical bundles in their written productions, despite being taught these expressions in reading materials. Similarly, Hyland (2008) identified significant differences between student and professional written performances regarding the structure and functions of lexical bundles, noting that "the research articles contained far fewer clusters and far fewer different clusters overall …. they revealed more participants strings and included a far higher proportion of text-oriented clusters" (Hyland, 2008, p. 59).

Other studies have investigated the usage of bundles between native and non-native writings (Chen & Baker, 2010; Ädel & Erman, 2012). These studies revealed that lexical bundles in non-native students' academic writing contained more VP-based bundles and discourse markers than in native academic writing, which appears to indicate "immature writing" (Chen & Baker, 2010, p. 44). In contrast, native academic writing included more NP-based and referential bundles. Moreover, non-native writers underused some high-frequency bundles found in native academic writing and overused certain lexical bundles that native writers rarely used.

In a study comparing non-native academic writing by L1 speakers of Swedish and undergraduate native speakers of linguistics, Adel and Erman (2012) concluded that non-native speakers use a limited and less varied range of lexical bundles including "unattended '*this'* constructions, existential '*there'* constructions, hedges and passive constructions" (Adel & Erman, 2012, p. 90).

Wei and Lei (2011) examined the use of lexical bundles in advanced Chinese EFL learners' academic writings and observed advanced learners used "similar number of prepositional phrases, noun phrases, be+ noun/ adjectival phrases and other structures of bundles as professional writers" (p. 164). Functionally, advanced Chinese EFL learners produced a similar amount of research-oriented and text-oriented bundles and fewer participant-oriented bundles compared to published writers.

Additionally, Karabacak and Qin (2013) compared the use of reference bundles in the argumentative essays of Turkish, Chinese, and American writers. They concluded that the majority of the bundles found in American writers' papers were absent from non-native speakers' essays due to differences in essay topics, lack of lexical bundle knowledge, and failure to produce correct bundles. In a similar vein, Muşlu (2018) investigated the stance lexical bundles in the argumentative essays by native English speakers and Turkish and Japanese EFL learners. It was concluded that native speakers use less but various lexical bundles. While lexical bundles have been used higher in Japanese and Turkish corpora, lexical variety is lower than the native data.

Uçar and Zarfsaz (2022) compared 80 argumentative essays written by Turkish EFL learners with 50 essays from native speakers in the British Academic Written English (BAWE) corpus. The analysis focused on three-word lexical bundles, with a frequency cut-off point of 20 occurrences in at least 5 different texts. The study found that Turkish students used a less diverse and more limited number of lexical bundles compared to their native English counterparts. English students employed a broader variety of structural types. Functionally, English students used more referential and stance bundles, while Turkish students relied more on discourse organizers (Uçar & Zarfsaz, 2022).

Methodology

Frequency Approach

This study employs a frequency-based (corpus-driven) approach, an inductive method as described by Sinclair (1987) and Nesselhauf (2004), to

extract lexical bundles from both corpora. Using frequency as the basis is an effective means of identifying usage patterns that "often go unnoticed by the researchers" (Biber et al., 2004, p. 376). As Tognini-Bonelli emphasizes (2001, p. 87) "linguistic categories are systematically derived from the recurrent patterns and the frequency distributions that emerge from language in context" within grammatical framework.

Research Design

The research design of this study is structured to provide a comprehensive analysis of lexical bundles in learner and native corpora. The steps of the data analysis followed in this study are presented in Table 3 below.

Table 3

Research Types and Stages of the Study

Stages	Process	Research Type
Stage 1	Automatic generation of frequency lists	Descriptive analysis
Stage 2	Selection of meaningful LBs manually	Corpus-based approach
Stage 3	Application of statistical analysis across corpora	Quantitative corpus analysis
Stage 4	Analysis of LBs structurally and functionally	Descriptive analysis
Stage 5	Comparison of LBs between 2 corpora	Interpretative analysis

The research design follows a systematic and structured approach. Initially, all the 3 and 4-word sequences were extracted from LOCNESS and TICLE using Antconc (version 3.5.9 for Windows) (Anthony, 2020). The sequences that appeared in at least three different texts were selected to ensure the accuracy of the identified bundles. This criterion helps in filtering out idiosyncratic or text-specific combinations, focusing instead on recurrent patterns.

Among these sequences, meaningful multiword items were manually identified. Overlapping bundles, as well as repetitive and erroneous ones, were removed from the list. This manual filtering ensures that only relevant and accurate lexical bundles are included in the analysis, enhancing the reliability of the findings. In the context of lexical bundle research, the issue of

overlapping bundles—where shorter expressions are embedded within longer ones—presents a significant challenge in both structural and functional analyses. The recent study by Cortes and Lake (2023) offers a solution to this problem by introducing the Lexical Bundle Identification and Analysis Program (LBiaP). This tool is specifically designed to identify and differentiate between overlapping bundles, ensuring that each identified bundle is treated as an independent observation. Their research emphasizes the importance of addressing complete overlapping, complete subsumption, and interlocking bundles, which are often overlooked in traditional analyses that focus solely on 3- or 4-word bundles. By accounting for these complexities, the LBiaP tool enhances the accuracy of lexical bundle categorization and classification.

Subsequently, the bundles in each corpus were categorized based on their structural and functional characteristics. The categorization process followed Biber et al.'s (1999; 2003; 2004) framework, which provides a comprehensive taxonomy for analyzing lexical bundles. This framework was chosen for its robustness and widespread acceptance in corpus linguistics research. The systematic categorization it provides enables a detailed and nuanced analysis of the structural and functional properties of lexical bundles. Additionally, the study examined the similarities and differences in frequent 3- to 4-word formulaic patterns between Turkish EFL learners and native speakers of English.

This comprehensive analysis allows for a detailed comparison and understanding of the use of lexical bundles in the written performances of both groups, identifying specific areas of divergence and convergence in the use of lexical bundles between the two groups. Thus, it aims to provide valuable insights into their writing proficiency and usage patterns.

Corpora

The Turkish International Corpus of Learner English (TICLE)

The Turkish International Corpus of Learner English (TICLE) was selected as the primary corpus for this study. TICLE is the Turkish sub-corpus of the International Corpus of Learner English (ICLE), designed by the Centre

for English Corpus Linguistics at the University of Louvain. TICLE consists of 280 essays totaling 199,532 words, all adhering to an argumentative style. Essays were collected from three Turkish universities, with an average word count of 713 per essay. The participants, predominantly female (81%) with an average age of 22.08, are university undergraduates still learning English as a foreign language. The representativeness of TICLE makes it an ideal corpus for analyzing the writing proficiency of Turkish EFL learners.

Louvain Corpus of Native English Essays (LOCNESS)

The Louvain Corpus of Native English Essays (LOCNESS) serves as the native speaker reference corpus, compiled by the Centre for English Corpus Linguistics at the Catholic University of Louvain. Covering the period from 1991 to 1995, the LOCNESS includes British and American university and A-level student essays, totaling 324,304 words. For this study, only the argumentative essays by American university students were used, encompassing 149,574 words from 175 essays. The age range of these participants (17–28 years) matches the TICLE participants, ensuring comparability. Topics in these essays cover a broad spectrum, including sex equality, water pollution, gender roles, and violence, among others. The comprehensive nature of LOCNESS ensures that it is a suitable parallel corpus for contrastive interlanguage analysis with TICLE.

Data Analysis Tools and Statistics Used

Cut-off Point

Formulaic sequences have to meet a set of criteria to be considered as bundles or multiword items such as identifying these items by corpus analysis tools (e.g., AntConc, Lancsbox) by using the criterion of frequency cut-off; and assigning functions to the sequences identified by frequency and range criteria.

The frequency cut-off point varies in each study. The actual frequency cut-offs in lexical bundle research vary from ten occurrences per million words (Biber et al., 1999; Biber, 2006), to forty occurrences (Biber, Conrad, & Cortes,

2004). Lower cut-offs may be used in smaller corpora (Biber et al., 1999) and determining the frequency cut-offs base on the earlier researchers rather than a statistical and empirical set of standards. Some other researchers such as Hyland (2008) employed a percentage criterion that required a sequence to appear in at least 10 percent of the texts in a corpus. Additionally, these MWIs have to appear in at least three to five different texts (Biber & Barbieri, 2007). This way, focusing on idiosyncratic uses by the authors is prevented. The cut-off point for the TICLE corpus was set as 10, while for LOCNESS it was set at 7. These cut-off points were determined to balance the need for capturing frequent patterns while avoiding noise from less common sequences. The cut-off points in the analysis of both corpora have been determined with the aim of increasing the accuracy and reliability. Also, the sequences which were seen in at least 3 different texts have been chosen during the identification of sequences in order to make the analysis more accurate.

Type/Token Ratio

In lexical frequency research, type-token ratio (T/t) is calculated in both native and learner corpora in order to reveal the lexical variety. T/t counts the number of different words in a text. Williamson (2013) emphasized that T/t ratio is a helpful measure of vocabulary variation in a written text or a speech. The T/t results are used to draw conclusions on lexical richness in learner texts (Granger, 2002) and it is computed by means of the following formula:

$$\text{T/t ratio} = \frac{\text{Number of word types} \times 100}{\text{Number of word tokens} \times 1}$$

T/t is obtained by dividing the type count by the token count, which is always <=1. While a high T/t score indicates a high degree of lexical variation, a low T/t indicates a low degree of lexical variation. T/t can also be expressed as a percentage, multiplying the ratio by 100. T/t ratio of MWIs represents the percentage of each item within all words in a corpus, which is the actual number of MWIs that fall into per 100 words. Tokens are the number of words

in a text. Yet, many of these tokens are repeated in corpora. In the study, the very same formula was utilized in the data analysis.

Log-likelihood

In corpus-based studies, frequency distribution of the corpora needs to be tested since the differences found between the frequencies of the items in corpora may be random or chance happening. In order to find out whether the differences are statistically significant, some tests are carried out. One of them is Log Likelihood (LL) analysis which helps us to normalize the sizes of the corpora we compare and reveal the significant differences between frequencies. LL may also be called as G-square or G-score. If results are significant, we are reasonably certain (usually 95% certain, sometimes 99% certain) that these results are not due to chance.

The observed and expected values are compared with LL in two datasets. While the observed values are the actual frequencies extracted from corpora, expected values are the frequencies that one would expect if no factor other than chance were affecting the values. The greater the difference between the observed and the expected values, the less likely it is that the difference has arose by chance. When the frequencies for corpus searches are obtained, the results are made comparable by converting frequencies to percentages or per million words. This process is referred to as normalizing the frequencies. However, normalized scores do not necessarily indicate statistical significance. Based on the provided information, data analysis was conducted using the Log-Likelihood (LL) method in the study.

Findings

Introduction

The overall descriptive results of the lexical bundles are presented in Table 4. The TICLE corpus consists of 280 texts with 199,532 words, while the LOCNESS corpus has 175 texts with 149,574 words. Turkish EFL learners

used 340 multiword items (MWIs) in their written productions, whereas native speakers used 349 MWIs. When examining their percentages, only 0.17% of Turkish learners' productions are considered bundles, compared to 0.23% of native speakers' production. The overall descriptive results indicate that while the overall number of lexical bundles used by Turkish EFL learners and native speakers is relatively close, the percentage of bundles in the total word count is higher for native speakers, suggesting a higher integration of multiword units in their writing. This aligns with Granger's (1998) findings that native speakers tend to use a wider variety of lexical bundles.

Table 4

The Bird-eye View of the Findings

Corpus texts	Total number of texts	Total number of words	Total number of LBs	% of LBs of
TICLE	280	199,532	340	0.17
LOCNESS	175	149,574	349	0.23
TOTAL	**455**	**349,106**	**689**	**0.40**

Lexical Diversity

To reveal the lexical diversity in each corpus, the type/token ratios of lexical bundles were calculated, as shown in Table 5 below.

Table 5

Type/token Ratio Results

	TICLE Type/Token	**LOCNESS Type/Token**	**Type/Token Ratio Results**
Verb Phrase Fragments	145/3466	105/1350	4.1% / 7.7%
Dependent Clause Fragments	84/1855	90/1132	4.5% / 7.9%
Noun Phrase and Prepositional Phrase Fragments	111/3174	154/2247	3.4% / 6.8%
TOTAL	**340/8495**	**349/4729**	**4% / 7.3%**

Note. T/t ratio= Type/token ratio; percentage of number of lexical bundles (types) in total of words (tokens) in each corpus.

The breakdown of each fragment type is presented in Table 5. The TICLE corpus has 340 types and 8495 tokens, while LOCNESS has 349 types and 4729 tokens. The type/token ratio reveals a significant

difference in lexical diversity between TICLE and LOCNESS. Native speakers exhibit a higher lexical diversity (7.3%) compared to Turkish EFL learners (4%). This supports the findings of Nesselhauf (2005) and Hyland (2008), which highlight the limited lexical range in non-native writing.

Research Question 1: What are the most frequent 3- and 4-word lexical bundles used in argumentative essays of Turkish EFL learners in TICLE and native speakers in LOCNESS corpora?

The first 100 most frequent LBs have been chosen in order to answer the first research question in each corpus, respectively. Out of 100 most frequent expressions, the following 40 were common in both corpora: *men and women, a lot of, it is not, in order to, on the other hand, they do not, the most important, in the world, there is no, most of the, there is a, it is a, do not have, it is the, they are not, that it is, there are many, because of the, to have a, should not be+V3, that they are, is not a, the right to, according to the, they want to, in the past, is one of the, to be a, in the future, look at the, this is a, as a result of, part of the, because they are, the fact that, all of the, you do not, do not want, this is not, the number of.* These bundles are used more frequently in the TICLE compared to the LOCNESS. The most common 10 bundles may be seen in the table below.

Table 6

The most Common 10 Bundles in both Corpora

TICLE		**LOCNESS**	
LBs	**Freq.**	**LBs**	**Freq.**
men and women	*126*	the death penalty	71
a lot of	*123*	*the fact that*	*71*
it is not	*120*	one of the	63
in order to	*119*	in the united states	53
on the other hand	*119*	*the right to*	*51*
they do not	*119*	***in order to***	50
the most important	*109*	the united states	48
in the world	*108*	*because of the*	*47*
there is no	*106*	*that it is*	*47*
most of the	*98*	the use of	45

As can be seen from the table, the only common bundle among the most frequent 10 bundles seems to be *in order to* which is in bold. What is obvious in the table is the frequency of the bundles in both groups. Tokens of the bundles in the TICLE are considerably more than the token in the LOCNESS. The most frequent bundles used by Turkish EFL learners include *men and women* (f: 126)*, a lot of* (f: 123), *it is not* (f: 120), *in order to* (f: 119), *on the other hand* (f: 119), *they do not* (f: 119), *the most important* (f: 109), *in the world* (f: 108), *there is no* (f: 106), *most of the* (f: 98). These ten most frequent bundles, totaling 1,147 occurrences, represent 13% of the total number of bundles within the TICLE corpus.

In contrast, in the LOCNESS corpus, the most frequent expressions in the LOCNESS corpus include *the death penalty* (f: 71)*, the fact that* (f: 71), *one of the* (f: 63), *in the United States* (f: 53) *the right to* (f: 51), *in order to* (f: 50), *because of the* (f: 47), *that it is* (f: 47), *the use of (*f: 45), and *it is a* (f: 40). These ten most frequent bundles occur 546 times, constituting 11% of all the tokens in the LOCNESS. Some of these expressions such as *in order to* (f: 119), *one of the* (f: 63), *the use of* (f: 45), *the fact that* (f: 71), *there is a* (f: 86) are among Biber et al.'s (1999) most common 3-word lexical bundles found in the expert academic writing. This suggests that native speakers use constructions more similar to expert academic writing than Turkish EFL learners.

The comparison of the most frequent lexical bundles indicates that Turkish EFL learners frequently use conversational-type bundles such as a lot of and men and women, which are less common in academic writing by native speakers. This is consistent with the observations of Chen and Baker (2010) and Bychkovska and Lee (2017), who noted the overuse of conversational bundles in learner writing.

The findings suggest that Turkish EFL learners overuse verb phrase fragments while underusing noun phrase and prepositional phrase fragments, in line with Biber et al. (1999). The presence of lexical bundles like in order to and the fact that in both corpora suggests some commonality in academic writing patterns, but the overall usage reflects the greater proficiency of native speakers in producing structurally varied bundles.

Research Question 2: How are these lexical sequences found in TICLE and LOCNESS corpora classified based on their structural and functional characteristics?

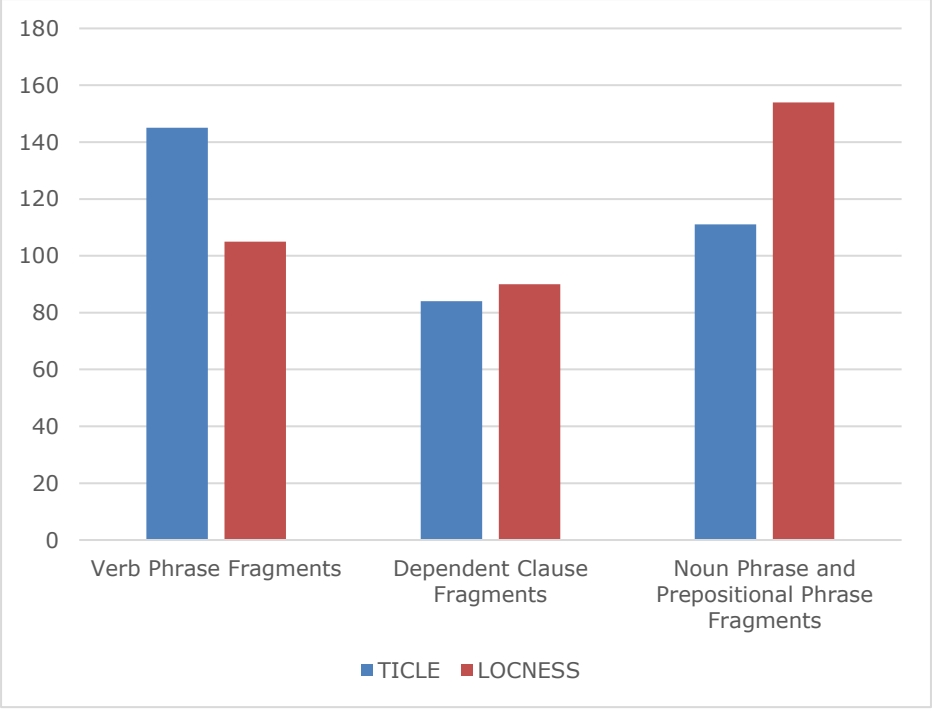
To address this research question, the two corpora were first compared based on their structural classification.

Comparison of the Structural Classifications in Two Corpora

Figure 1 indicates the distribution of the structural types of both corpora.

Figure 1

Structural Classifications in Both Corpora



Analysis of the structures used in both corpora reveals that while Turkish EFL learners predominantly use verb phrase (VP) fragments, native speaker students frequently use noun phrase (NP) and prepositional phrase

(PP) fragments. NP and PP fragments are the second most frequent category used in TICLE, constituting 33% of the classification, whereas VP fragments are the second most frequent structures in LOCNESS, constituting 33% of all the bundles. Dependent clause fragments are the least frequently used fragments in both corpora.

The findings align with Biber et al. (1999) and Granger (1998) in highlighting that Turkish EFL learners predominantly use VP fragments, which is characteristic of learner language. The underuse of NP and PP fragments suggests a limited structural range in the learners' writing.

When examining the sub-categories of these fragments in detail, personal pronouns+VP fragments is the only category constituting 42.4% of all structures in TICLE, these structures constitute only 32% in LOCNESS. Two examples are provided below:

(12) In conclusion, **I can say that** we cannot limit education of [TRME3026]
(62) The only problem is that **they do not** give any data that provides [USARG_0047]

Within the VP fragments, a significant number of phrases produced by Turkish EFL learners report negative states, as illustrated by examples such as *students/universities/people/I/you/they/we/it do(es) not* (f: 214), *they/we/people are not* (f: 90), *they/it should not be* (f: 25), *they should not* (f: 11), *I/they/you/we/he/women cannot* (f: 194)…etc. These negative structures constitute 22.6% of the personal pronouns + VP fragments category. This pattern is not observed in the native speakers' usage.

(10) **Women cannot** make a stand against men. [TRKE2025]
(14) is not an help to their friends and **they should not** consider it as a way to [TRKE2026]

The frequent use of negative structures in TICLE suggests a limited variety in expressing negation, aligning with findings from Nesselhauf (2005) and Hyland (2008). This repetitive usage underscores the need for teaching more varied and nuanced expressions of negation.

Additionally, it was revealed in TICLE that many of the fragments within the personal pronouns + VP fragments include bundles with embedded *be*-verbs. Turkish EFL learners also use these existential *there*-constructions, as shown in examples like *there are some/many/a lot/lots of/many people/ also/so many* (f: 215), *there is a/an* (f: 108)*, there is no* (f: 89), *there is no need* (f: 17), *there is not* (f: 26), *there is nothing* (f: 20). Examples include:

(7) we can realize that **there are lots of** things became a part of [TRCU1018]
(8) So, **there is no need** for male dominance. [TRCU1115]
(9) **there is a** balance between school community and individual [TRKE2011]

The patterns observed in TICLE are not mirrored in LOCNESS. The analysis revealed that non-native speakers used a significantly higher percentage of existential *there* bundles than their native speaker counterparts. Additionally, the research found that Turkish L1 learners used more evaluative bundles, such as *anticipatory it patterns* like *it is easy to, it is a fact*… etc. than native speakers.

The overuse of existential *there* constructions and *be-verbs* in TICLE indicates a reliance on simpler grammatical structures, a finding consistent with Chen and Baker (2010). This overreliance may point to gaps in learners' grammatical competence, highlighting areas for targeted pedagogical intervention.

Another frequent usage in VP fragments is the structures with modals used by Turkish EFL learners, constittuting 35.1% of all VP fragments. Examples include; *they should not, I can say that, you can see, it can be*…etc.

(19) To all human life **should be given** equal protection under the law. [TRCU1091]
(21) Cheating **cannot be annihilated** completely. [TRCU1162]
(22) Euthanasia **should not be allowed** since it has many objections [TRCU1140]

The preference for evaluative bundles and modals reflects the learners' attempt to convey stance and modality, albeit repetitively. This echoes findings

by Bychkovska and Lee (2017) and suggests the need for more explicit instruction on diverse evaluative expressions and modal usage.

Turkish EFL learners did not produce *"discourse marker + VP fragment"* bundles such as *I mean you, you know it was and I mean I don't* and *"WH-question fragments"* including *what do you think, how many of you, what does that mean…etc*. Similarly, NS did not used *"discourse marker + VP fragment"* or *Yes/no question fragments,* and they used only one wh-question fragment in the entire dataset. In this regard, the usage patterns are similar.

Another notable finding is that while Turkish EFL learners frequently use *who*-relative clauses with VP bundles, and LBs containing embedded *who*-clauses such as *people/students/women who are, the people who are, people who have,* these structures are not found in the native corpus.

(26) Today there are a lot of successful **women who are** leading whole companies and even whole [TRCU1035]

The frequent use of who-relative clauses in TICLE, absent in LOCNESS, indicates learners' preference for certain syntactic structures possibly due to L1 transfer, as discussed in Granger and Paquot (2008). This finding highlights the importance of addressing relative clause usage in EFL instruction.

Additionally, another significant finding is the underuse (35.64- LL value) of *to-clause fragments* (copula be+adjective/noun phrase) within the dependent clause fragments by Turkish EFL learners in comparison to the native speakers. And lastly, the noun phrase with of-phrase fragments have significantly underused in TICLE with 40.54- LL ratio. Some of these phrases used by native speakers are *the number of, the amount of, the use of, because of the, one of the, invention of the, the idea of, out of the, this type of*. These bundles were not observed in TICLE.

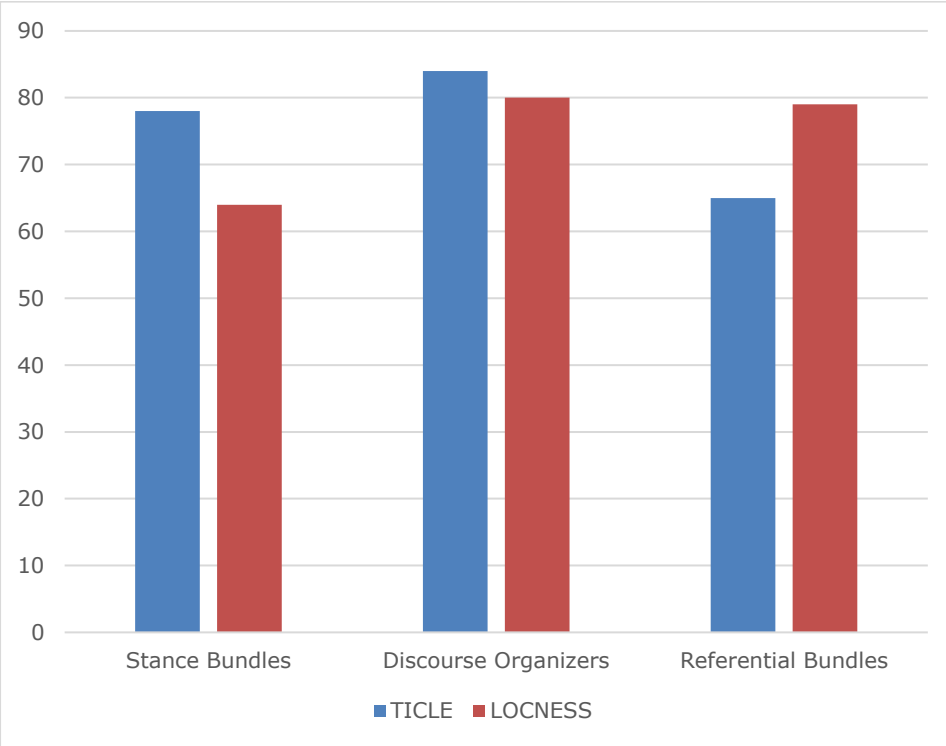
Comparison of Functional Classifications in Two Corpora

When comparing the productions in terms of functional properties in discourse, there is not much difference between the two corpora. The functional classification revealed that 67% of bundles in TICLE and 64% of the bundles

in LOCNESS did not have any functional properties. This significant portion indicates a dominance of structural rather than functional roles within the texts. This aligns with Granger and Paquot (2008), which highlight the need for more nuanced use of functional language in EFL learners. The classifications of both corpora may be seen in Figure 2.

Figure 2

Functional Classifications in Both Corpora



Both Turkish EFL learners (36%) and native speakers (35%) predominantly used discourse organizers. Some of the most common examples include *on the other hand* (f: 119), *as you see* (f: 86), *in order to* (50), *that it is* (47), *that they are* (33). These discourse organizers help in structuring the text, guiding the reader through arguments and discussions. The high use of discourse organizers by both Turkish EFL learners and native speakers indicates their crucial role in structuring argumentative essays. This

corroborates findings from Hyland (2008) and Chen and Baker (2010), emphasizing the importance of teaching effective use of discourse organizers to improve coherence and cohesion in learner writing.

(135) **If we look at** this subject from another aspect, [TRCU1038]
(136) Secondly, **when we think** about psychological situation and future life [TRCU1158]
(165) twenty one are considered to be minors, **when it comes** to drinking alcoholic beverages in [USARG_0163]

Stance bundles were the second most frequently used function in TICLE (35%), while referential bundles were the second most common function in LOCNESS (34%). The preference for stance bundles by Turkish EFL learners and referential bundles by native speakers suggests differing strategies in academic writing. This indicates that Turkish EFL learners often use stance bundles to express opinions, beliefs, and attitudes, while native speakers more frequently use referential bundles to provide context, specify attributes, and refer to time, place, or text. This finding supports Biber et al. (2004) and Ädel and Erman (2012), which point to the importance of teaching a balanced use of stance and referential bundles to enhance argumentation and description in EFL writing.

Some of the stance bundles in TICLE include *I believe that* (f: 43), in my opinion (f: 70), *I/they/you/he want(s) to* (f: 139), *the fact that* (f: 31). Example sentences include:

(121) **I believe that** euthanasia should be legalized albeit be [TRCU1176]
(122) about the capital punishment should be **the fact that** murder is a crime punishable by [TRCU1103]

Some of the most common referential bundles in LOCNESS include *at the end of* (f: 9), *at the beginning of* (f: 8), *the most important* (f: 18), *the number of* (f: 26), *a lot of* (f: 29). Example concordance lines from LOCNESS are:

(173) This would save the government **a lot of** money and make people support [USARG_0152]
(180) They do this because **at the beginning of** the next fiscal year [USARG_0146]

Referential bundles were the least common category used by Turkish EFL learners (29%). Examples of referential bundles from TICLE include *the people who* (f: 67), *this is a/the* (f: 60), *in the past* (f: 41), *in the world* (f: 108), *in terms of* (f: 18). These bundles often refer to people, time periods, or broad concepts, indicating a reliance on general terms rather than specific, context-driven references. The absence of imprecision bundles *like* or *something like that* in both corpora indicates an area where learners may benefit from explicit instruction to enrich their use of vague language for generalization and hedging. Example sentences are:

(145) have to be dependent especially **in terms of** money on their family. [TRME3001]
(146) die can determine to apply to euthanasia as **the result of** a violent pain or a sudden [TRCU1076]

In contrast, stance bundles were the least frequently used function in LOCNESS (28%). Examples include *the fact that* (f: 71), *I would like to* (f: 7), *I believe that* (f: 12), *I think it* (f: 7), *I think that* (f: 9). These bundles allow writers to assert their viewpoints or present subjective statements, which are less prevalent in native speakers' academic writing, potentially due to a more objective and detached writing style. Example concordance lines are:

(153) often stand by its use based on **the fact that** it is provided for in [USARG_0012]
(154) Life is considered unquestionable, but **I would like to** give it more scrutiny. [USARG_0037]

Regarding the referential bundles, these are the least frequently used expressions by Turkish EFL learners, while native speakers used them as

the second most frequent bundles. Within the referential bundles, one sub-categories is imprecision which include bundles like *or something like that.* These were produced in neither TICLE nor LOCNESS, indicating a potential area of underuse that could be targeted in instructional materials.

Research Question 3: How similar or different are the frequently used 3- to 4- lexical sequences used by Turkish EFL learners from those found in native English speakers' (LOCNESS) corpora in terms of structure, function, and diversity in their use?

To determine the similarities and differences between the frequently used lexical sequences in TICLE and LOCNESS, a log-likelihood analysis was conducted. The analysis focused on three aspects: structure, function, and diversity of use.

a. Log-likelihood Results of Structural Classifications in Both Corpora

The structural classifications of lexical bundles in both corpora are presented in Table 7 below.

Table 7

Log-Likelihood Results of Structural Classifications in Both Corpora

Structures	Freq. in Corpus TICLE	Freq. in Corpus LOCNESS	Log-likelihood	Sig.
Verb Phrase Fragments	145	105	0.07	0.787 +
Dependent Clause Fragments	84	90	5.54	0.019*-
Noun Phrase and Prepositional Phrase Fragments	111	154	24.88	0.000***-
TOTAL	340	349	16.97	0.000***-

Note. **+ indicates overuse** in TICLE relative to LOCNESS, **- indicates underuse** in TICLE relative to LOCNESS

The log-likelihood analysis reveals that all structural categories, except for verb phrase (VP) fragments, were significantly underused in TICLE.

The total LL value of 16.97 indicates a significant underuse overall. The most notable difference was observed in noun phrase (NP) and prepositional phrase (PP) fragments, with an LL value of 24.88-, suggesting a marked underuse by Turkish EFL learners. Dependent clause fragments also showed a significant underuse, with an LL value of 5.54-. The insignificant difference in VP fragments indicates similar usage frequencies between the two corpora. This underuse of types highlights a lower lexical variety and richness among Turkish EFL learners compared to native speakers. The total number of tokens in both corpora is shown in the following table.

The insignificant difference in the usage of VP fragments suggests a similar preference for these structures in both corpora. This aligns with findings by Biber et al. (2004) and Chen and Baker (2010), indicating that learners and native speakers both heavily rely on verb phrase fragments in their writing. Further, Granger's (1989) work highlights the importance of these fragments in learner language, often as formulaic expressions used for syntactic simplicity.

The significant underuse of NP and PP fragments by Turkish EFL learners suggests a need for focused instructional strategies to enhance their usage. According to Granger (1998), native-like proficiency in academic writing involves a substantial use of NP and PP structures, which contribute to syntactic complexity and lexical richness.

Table 8

Log-Likelihood Results of Structural Tokens of LBs in Both Corpora

Structure	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
Verb phrase fragments	3466	1350	452.05	0.000***+
Dependent clause fragments	1855	1132	30.20	0.000***+
Noun phrase and prepositional phrase fragments	3174	2247	4.32	0.038 *+
TOTAL	8495	4729	275.81	0.000**+

Note. **+ indicates overuse** in TICLE relative to LOCNESS, **- indicates underuse** in TICLE relative to LOCNESS

The number of tokens in each structural category shows a significant overuse in TICLE across all categories. The VP fragments exhibit the most considerable overuse, with an LL value of 452.05+, far exceeding the values of other categories. Dependent clause fragments also show significant overuse, with an LL value of 30.20+. Even though the variety of bundles (types) is limited in TICLE, the repetitive use of a few bundles (tokens) is markedly higher. This indicates that Turkish EFL learners rely heavily on a restricted set of bundles, using them repeatedly in their writing.

The repetitive use of a limited number of bundles in TICLE highlights a potential area for pedagogical intervention. Encouraging learners to expand their repertoire of lexical bundles through exposure to diverse academic texts and targeted exercises could foster greater lexical variety.

b. Log-likelihood Results of Functional Classifications in Both Corpora

The log-likelihood (LL) ratio of functional types in TICLE and LOCNESS is presented in Table 9 below.

Table 9

Log-Likelihood Results of Functional Types in Both Corpora

Functional Structures	Freq. in Corpus TICLE	Freq. in Corpus LOCNESS	Log-likelihood	Sig.
Stance bundles	78	64	0.29	0.593 -
Discourse organizers	84	80	2.34	0.126 -
Referential bundles	65	79	8.38	0.004 ** -
TOTAL	227	223	8.19	0.004 ** -

Note. **+ indicates overuse** in TICLE relative to LOCNESS, - **indicates underuse** in TICLE relative to LOCNESS

According to the functional classification, all of the categories have been underused by Turkish EFL learners in terms of the LL measurement of types. This suggests that most of the structures used by Turkish EFL learners do not have functional properties. Two of these structures have been

significantly underused. The most significant difference is in the referential bundles category, with an LL value of 8.38-, indicating a significant underuse compared to native speakers. The second most significant difference is in discourse organizers, with an LL value of 2.34-, also underused in TICLE. Other differences are not considered significant according to LL measurement ($p < 0.05$). There is a significant underuse of functional properties in TICLE compared to LOCNESS. The total difference is 8.19-, indicating that lexical variety and richness in TICLE are less than in native speakers in terms of using functional bundles.

The LL ratio of tokens of these functional categories is shown in Table 10.

Table 10

Log-Likelihood Results of Functional Tokens in Both Corpora

Functional Structures	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
Stance bundles	1768	804	145.70	0.000 ***+
Discourse organizers	2182	965	197.52	0.000 ***+
Referential bundles	1775	1167	12.22	0.000 ***+
TOTAL	5725	2936	289.77	0.000 ***+

Note. **+ indicates overuse** in TICLE relative to LOCNESS, **- indicates underuse** in TICLE relative to LOCNESS

When examining the LL ratio of tokens in the functional classification, all categories have been extremely overused by Turkish EFL learners, with a total LL value of 289.77+. This indicates that even though all types of these categories have been significantly overused in TICLE, the number of tokens shows an extreme overuse. In other words, the same structures have been produced numerous times in TICLE.

Discourse organizers are the most overused structures among functional structures with an LL value of 197.52+, representing the most significant difference in the entire functional classification. The significant overuse of discourse organizers in terms of tokens suggests that Turkish EFL

learners frequently use these bundles to manage discourse flow. However, the underuse in type frequency indicates limited variety. This finding is consistent with studies by Chen and Baker (2010) and Adel and Erman (2012), who noted similar trends among EFL learners.

Stance bundles are the second most significant difference in comparison to LOCNESS. While Turkish EFL learners underuse stance bundles in terms of types, the token frequency indicates a significant overuse. This suggests a reliance on a few familiar bundles, possibly due to a lack of exposure to a broader range of stance expressions. Aligning with Granger and Paquot (2009), it is essential to encourage learners to diversify their use of stance bundles to enhance their argumentative writing.

The least significant difference belongs to referential bundles with an LL value of 12.22+. The underuse of referential bundles in terms of types and their overuse in tokens highlights a gap in learners' ability to use these bundles effectively. This aligns with Hyland (2008) and Shin's (2018) findings, emphasizing the need for targeted instruction on referential expressions to improve lexical diversity and precision.

Discussion

Discussion for Research Question 1

The underuse of certain bundles (e.g., to-clause fragments, of-phrase fragments) by Turkish EFL learners compared to native speakers aligns with the findings from previous studies, which assert that novice writers infrequently use academic-register lexical bundles. Chen and Baker’s (2010) study supports this, showing that native speakers exhibit a wider diversity in their use of bundles compared to their non-native counterparts, a trend also reported by Bychkovska and Lee (2017).

The presence of 40 out of 100 common bundles in both corpora is consistent with Adel and Erman's (2012) research, which identified 60 shared bundles among a total of 130 within their dataset. Conversely, these results contrast with Chen and Baker's (2010) study, which reported a higher shared bundle rate, with 54 shared instances out of a total of 78 bundles. They also

noted that ‘‘the use of lexical bundles in nonnative and native student essays is surprisingly similar’’ (p. 44).

Examining the most frequent items in TICLE, bundles such as *a lot of, the most important, in the world* are considered to be conversation-type bundles, commonly used by learners. This finding is consistent with research by Staples et al. (2013), Chen and Baker (2010), and Bychkovska and Lee (2017). The frequent use of these types of bundles is often cited as a distinctive feature of learner writing. Additionally, Shin (2018) observed a significant presence of these types of lexical bundles in her study, suggesting that their usage is not confined to a specific learner population but rather indicative of novice academic writing, irrespective of the first language.

Furthermore, the structures of the bundles extracted from both TICLE and LOCNESS are not complete grammatical units, as illustrated in the examples provided. This finding is consistent with Biber et al.’s (1999) study, which found that more than 95% of lexical bundles were not complete units in academic writing. Cortes (2004) supports this argument, stating that “lexical bundles are identified empirically, rather than intuitively, as word combinations that recur most commonly in a register, and therefore, lexical bundles are usually not complete structural units, but rather fragmented phrases or clauses with new fragments embedded” (p. 400).

Granger (1998) highlights the role of cross-linguistic influence and transfer in the use of lexical bundles by non-native speakers, which might explain the frequent use of certain conversational bundles in TICLE. Additionally, Paquot (2010) emphasizes the importance of discipline-specific corpora in identifying the functions of lexical bundles, which can provide a more nuanced understanding of their use in academic writing.

Discussion for Research Question 2

Structural Characteristics

Lexical bundles in both the TICLE and LOCNESS corpora were classified according to their structures and functions using Biber et al.’s (2004) taxonomy. Out of 340 distinct bundles, the initial category of the structural

classification pertains to verb phrase fragments, represented by 145 types and 3466 tokens, comprising 42.4% of all structures in the TICLE corpus. A comparison of the structures used in both corpora reveals that while Turkish EFL learners predominantly employed *verb phrase fragments*, native students leaned more towards *noun phrase and prepositional phrase fragments*. This implies that Turkish EFL learners produce more VP-based bundles relative to native speakers, corroborating the findings of Chen and Baker (2010) and Shin (2018), who highlighted that student writings typically contain more *VP*-based bundles than those written by native speakers. This trend is especially evident in specific subcategories.

Our categorization scheme also identified this trend in the VP fragments subcategory of personal pronouns + VP fragments, which constituted a quarter of all entries in the structural classification. This finding aligns with studies by Bal Gezegin (2019) and Wei and Lei (2011), which revealed a strong preference for noun phrase structures among native speakers.

Additionally, the analysis uncovered that a substantial portion of the fragments in this subcategory include bundles incorporating embedded *be-verbs,* with 44 types constituting 38.2% of all verb phrase fragments. This finding aligns with Chen and Baker (2010), who reported that a third of the LBs incorporated *be*-verbs, rendering student writing "simplistic and verbose" (p. 866). The authors posited that this overuse might be attributed to the learners' extensive reliance on existential *there*-constructions. This pattern was markedly evident in the argumentative essays of Turkish EFL learners. Examples from our study include *there are some/many/a lot/lots of/many people/ also/so many* (f: 215), *there is a/an* (f: 108), *there is no (f: 89), there is no need* (f: 17), *there is not* (f: 26), *there is nothing* (f: 20).

Contrary to these findings, Ädel and Erman (2012) indicated that native students used more existential-*there* constructions and passives, while non-native students tended to initiate arguments with evaluative bundles like *anticipatory it* patterns (e.g., *it is easy to*). Our research, however, discovered that non-native speakers used more existential *there*-bundles at a significantly higher rate than their native counterparts.

Within the verb phrase fragments, *discourse markers+VP fragments, wh-question* and *yes/no question fragments* were absent in both corpora. This absence could be attributed to the types of corpora studied, as both consist of written registers made up of argumentative essays. Biber et al. (1999) noted that such expressions are more common in spoken registers. Notably, these two structure types were also absent in Elturki's (2015) study, which examined the development of LBs among learners at various proficiency levels over a one-year period.

Another noteworthy finding pertains to *dependent clause fragments*, where learners frequently use *who-relative clauses* with VP bundles, and LBs containing embedded who-clauses such as *people/students/women who are, the people who are, people who have*. These structures do not appear in the native speaker corpus. Moreover, most of these structures involve *people*, indicating an overuse of this somewhat vague word, a pattern characteristic of learner writing. Many of the LBs produced include or collocate with *people*, aligning with previous studies of Chen and Baker (2010) and Bychkovska and Lee (2017).

Within the verb phrase fragments category, a substantial proportion of the expressions produced by Turkish EFL learners encapsulate negative states. This trend is exemplified by phrases such as *students /universities /people /I /you /they /we /it do(es) not* (f: 214), *they/we/people are not* (f: 90), *they/it should not be* (f: 25), *they should not* (f: 11), *I/they/you/we/he/women cannot* (f: 194) …, etc. These negative structures account for 22.6% of the personal pronouns +VP fragments category. This observation stands in contrast with the findings of Shin (2018), who noted that the native writers exhibited a preference for negatively phrased expressions such as *disagree with the statement* and *do not agree with*. Conversely, learner writers tended to favor positively phrased expressions like *agree with the statement* or *so I agree with*, whereas native speakers address negative aspects.

The finding of underusing *to-clause* fragments by Turkish EFL learners diverges from Chen and Baker's (2010) earlier study, which found that both L1 Chinese student writers and NS used "to-clause fragments" extensively, showing a preference for the frame "in order to + Verb."

A quite significant underuse of noun phrase with of-phrase fragments in TICLE in relative to LOCNESS is in line with Shin's (2018) study which reported that the native writers used significantly more noun phrases with of-phrase fragments than their learner counterparts.

Functional Characteristics

The functional analysis revealed that not all the used structures serve a function in the discourse, 67% of bundles in TICLE and 64% of the bundles in LOCNESS did not have any specific functions. The nature of argumentative essays, the specific genre used in our study, could explain this outcome since the primary objective of such essays is "to express [one's] opinion about an issue" (Staples et al., 2013, p. 217).

Discourse organizers were most extensively used by both Turkish EFL learners (36%) and native speakers (35%). This aligns with findings from Chen and Baker's (2010) study, which concluded that both British and Chinese students also employed a significant number of discourse organizers in their writing, particularly to elaborate and/or clarify a topic. Most of the structures used are verb phrase-based bundles, such as *'this means that', 'that is to say', 'can be used'*, etc. Staples et al. (2013) also revealed that in their corpus of EAP texts, more than half of the LBs used functioned as discourse organizers.

In the TICLE corpus, stance bundles were the second most commonly used function (35%), while in the LOCNESS corpus, referential bundles filled this spot (34%). Referential bundles were the least used by Turkish EFL learners (29%), whereas stance bundles were the least frequently used by native speakers (28%). There is a slight underuse of the functional properties in the TICLE corpus compared to the LOCNESS corpus. However, the overall difference between the two corpora in terms of functional properties is significant with a LL value of 8.19-.

Referential bundles, the least frequently used expressions by Turkish EFL learners, were used as the second most frequent bundles by native speakers. Adel and Erman (2012) revealed that the largest proportion of LBs functioned as referential expressions in their study on LBs in academic writing,

attributing the differences between referential expressions and discourse organizers to the frequent use of prepositional phrase-bundles. The minimal use of referential bundles is also observed in previous studies by Shin (2018), Chen and Baker (2010), and Salazar (2014), which found that both native and non-native novice writers used only a few reference bundles, with natives using slightly more than non-natives. Chen and Baker (2010) also noted that using fewer referential bundles is characteristic of novice writers regardless of their first language. In all these studies, the variations in the proportions of bundles in terms of their functions might indicate the different features of different genres. In other words, genre plays an important role in the use of bundles.

The observed differences in structural and functional classifications could be attributed to several factors. First, our non-native speakers are learning English as a foreign language, whereas some studies, including Chen and Baker's (2010) study involving Chinese L1 students, explore contexts where English is a second language. Greater disparities are anticipated when comparing groups from EFL and ESL contexts.

Another factor could be the methodology itself, as ''larger corpora will generate fewer recurrent word combinations with the same cut-off normalized frequency, when compared with smaller corpora, because large corpora will elicit higher converted raw frequencies'' (Chen & Baker, 2010, p. 43; see also Biber & Barbieri, 2007, p. 269). While we employed different cut-off points for each corpus, smaller cut-off points might have revealed different bundle patterns. The differences may also stem from the nature and impact of the (non-disciplinary) argumentative essays, which require students to express their own ideas and opinions and often embody characteristics of spoken genres, such as personal declarations. Besides, the essays were compiled under exam situations in both corpora. The differences with other studies might also stem from learners' reliance on their native language (L1 transfer), where they "choose words and phrases closely resembling their first language or those learnt early or widely used" (Hasselgren, 1994, p. 237). Furthermore, cultural writing styles or habits could also influence the writers.

The similarities between the two corpora in terms of functional properties are not particularly substantial. This could be attributed to the nature of the (non-disciplinary) argumentative essays,

Finally, it should also be noted that we excluded the erroneous and repetitive structures from our analysis. We did not explore the impact of errors on learners' lexical bundle usage, a common phenomenon in learner writing. For instance, Karabacak and Qin's (2013) comparison of lexical bundles used by Turkish, Chinese, and American university students highlighted that the non-native students might possess partial knowledge of a bundle and attempt to produce it, though often unsuccessful. This finding indicates that the learners either lack complete knowledge of the bundles or are unfamiliar with them altogether. Therefore, had we included erroneous and repetitive structures in our analysis, the results might have differed, as learners' attempts to produce lexical bundles could go unnoticed in automatic data-driven and frequency-based approaches (Shin et al., 2018).

Discussion for Research Question 3

Log-likelihood findings of the structural classifications revealed that dependent clause and NP fragments are underused in TICLE, while VP fragments reveal an insignificant overuse in terms of type frequency. This underuse points to a deficit in lexical variety and richness compared to native speaker language use. Nevertheless, the LL values corresponding to the tokens of these structures indicate an overarching overuse by Turkish EFL learners across all three categories. The overuse of these bundles may be attributed to learners' tendency to opt for the safest lexical options labeled as "lexical teddy-bear tendency" by Hasselgren (1994). As a result, Turkish EFL learners underuse other potential alternatives for these existing bundles by clinging to a limited set of phraseological *teddy bears*.

Within the structural classification, the sub-category *verb phrases with passive verbs*, there were no significant differences between two corpora, despite a minor underuse (LL value 1.21-) in TICLE compared to LOCNESS. This aligns with findings from Ädel and Erman (2012), Wei and Lei (2011) and Chen and Baker (2010), who found that native speakers used passive structures more than EFL learners and L1 Chinese students.

Another significant discrepancy is observed in terms of *to-clause fragments,* which Turkish EFL learners underused substantially. This finding

diverges from Chen and Baker's (2010) earlier study, which found that both L1 Chinese student writers and NS used "to-clause fragments" extensively, showing a preference for the frame "in order to + Verb." Additionally, a significant underuse of noun phrase with *of-phrase* fragments in TICLE is observed relative to LOCNESS. This result aligns with Shin's (2018) study, which found that the native writers used significantly more noun phrases with *of-phrase* fragments than their learner counterparts.

An interesting finding is the overuse of *prepositional phrases* by Turkish EFL learners both in terms of type and token LL values. The common expression *in the world* used 19 times by native speakers, while Turkish EFL learners used this expression 109 times. This subcategory features expressions specific to Turkish L1 learners such as *all over the world* (f: 32), *around the world* (f: 12), *all around the world* (f: 18), *in real life* (f: 16), *in the real world* (f: 16), and *for real world* (f: 16). Learners often over-generalize the use of all over the world, classifying it as a 'learner bundle' frequently used in L2 academic writing but rarely in native English academic writing (Chen & Baker, 2010). However, this finding contradicts Shin's (2018) study, which found that native speakers used many idiomatic expressions such as *in the long run* (20) and *in the real world* (17) more than the learners.

A review of token counts reveals a statistically significant overuse of nearly all categories in the functional taxonomy by Turkish EFL learners. The higher number of tokens in TICLE can largely be attributed to the repetitive overuse of a few specific bundles, a finding that aligns with Salazar's (2014) study. The overuse of bundle tokens by non-native speakers is also observed in Wei and Lei's (2011) and Jalali et al.'s (2008) studies. This finding also aligns with Chen and Baker's (2010) research, which found that native speakers produced the broadest variety of lexical bundles, while non-native speakers employed the narrowest range.

Beginning with the stance bundles used by both groups, the log-likelihood value shows an underuse by Turkish EFL learners in terms of types, but a significant overuse in terms of the tokens. This somewhat contradicts with prior research indicating that native writers use proportionally more stance bundles than learners in terms of both types and tokens, as suggested by Ping (2009) and Bychkovska and Lee (2017). Overall, the usage percentages of

epistemic devices by native speakers and Turkish EFL learners are fairly similar, with native speakers employing epistemic bundles slightly more

Within the functional classification, the most significant difference lies in the underuse of *time/place/text reference* bundles by non-native speakers. This could be due to their avoidance of using prepositions for time, place, and text references. Morris and Cobb (2004) argue that L2 learners often resort to avoidance strategies when dealing with multi-word items. It has also been highlighted in the literature that Turkish EFL learners may avoid using certain prepositions when using multi-word items.

Conclusion

This study has revealed significant differences in the use of lexical bundles between Turkish EFL learners and native English speakers, highlighting areas where Turkish learners exhibit less variety and overuse certain bundles compared to their native counterparts.

The analysis shows that Turkish EFL learners predominantly use verb phrase (VP) fragments, accounting for 42% of structures, while noun phrase (NP) and prepositional phrase (PP) fragments are more common among native speakers, making up 46% of their structural classification. In TICLE, NP and PP fragments are the second most common category, comprising 33% of the classification, whereas in LOCNESS, VP fragments represent 33% of all bundles. Dependent clause fragments are the least common in both corpora. This limited use may be due to native speakers producing structures by reducing relative clauses in their essays, which were not apparent in the analysis. For Turkish EFL learners, avoidance of these expressions, particularly those with relative clauses, is likely due to their intermediate proficiency level.

The repetitive use of certain bundles by Turkish EFL learners suggests a reliance on familiar structures, which could be due to concerns about making errors or a preference for using known expressions. This pattern aligns with the "lexical teddy-bear tendency" (Hasselgren, 1994), where learners cling to a limited set of safe, familiar bundles. The overuse of certain bundles may

indicate a lack of exposure to a wider range of lexical items and insufficient practice with varied expressions.

The analysis found that many bundles in TICLE include embedded *be*-verbs and existential *there* constructions. This reliance on simple and familiar structures can render student writing "simplistic and verbose" (Chen & Baker, 2010, p. 866). The overuse of be-verbs and existential constructions like there is/are (e.g., there are some/many/a lot/lots of/many people) is markedly higher among Turkish EFL learners compared to native speakers.

Significant underuse of to-clause fragments and noun phrase with of-phrase fragments was observed among Turkish EFL learners. This finding contrasts with studies indicating a preference for these structures in native writing. The underuse of these fragments may reflect a gap in learners' proficiency and familiarity with more complex grammatical structures.

The functional analysis revealed that not all used structures serve a function in the discourse, with 67% of bundles in TICLE and 64% in LOCNESS lacking functional properties. This outcome can be attributed to the nature of argumentative essays, whose primary objective is "to express [one's] opinion about an issue" (Staples et al., 2013, p. 217). Both Turkish EFL learners (36%) and native speakers (35%) predominantly use discourse organizers, with stance bundles being the second most common in TICLE (35%) and referential bundles in LOCNESS (34%). Referential bundles were the least used by Turkish EFL learners (29%), while stance bundles were the least frequently used by native speakers (28%). Despite a slight underuse of functional properties in TICLE compared to LOCNESS, the overall difference is significant, with an LL value of 8.19-. This suggests that the lexical diversity and richness in TICLE are less than in native speakers in terms of functional bundles.

The study concludes that Turkish EFL students exhibit a restricted range of lexical bundles in their academic writing, relying on a limited set of repeated patterns compared to native English speakers. This may be attributed to limited exposure of the bundles and first language influence. Underusing certain structures in the study may also be attributed to the avoidance strategies learners apply. The study emphasizes the need for explicit instruction on lexical bundles in EFL curricula to enhance the academic writing skills of Turkish students. In conclusion, this study highlights the need for

a more varied and contextually appropriate use of lexical bundles among Turkish EFL learners. By addressing these areas, educators can help learners develop greater lexical diversity and proficiency in academic writing, aligning more closely with native speaker norms.

Pedagogical Implications

The findings of this study provide valuable pedagogical insights for English Language Teaching as a Foreign Language. The lack of knowledge of bundles and overuse and underuse of the certain bundles implies some underlying problems in EFL teaching pedagogy in Turkish context. The learner corpus-based studies on the use of multiword items, as in this particular study, contributes to building learner profiles of the learners coming from various L1 backgrounds. The profiles of these learners offer significant insights that can guide the development of customized language teaching materials in ELT and inspire EFL teachers to design bespoke educational resources and corrective teaching activities, techniques, and methodologies. Some of the pedagogical implications are as follows:

Fostering learner awareness*:* Raising learners' awareness about the use of multiword items in their writings could serve as another effective strategy in teaching these constructs. The use of learner data in the classroom exercises such as comparing learner and native speaker data and analyzing errors in learner language has been suggested to raise awareness. To exemplify, one of the ways is *chunking of text* (Wood, 2015). In this teaching technique, sequences are highlighted and checked by means of online/offline tools or sources and learners begin to use more multi-word items in their writings. Also, activities such *as eliciting the collocations, completing collocations from memory, matching the words together, selecting the missing word* to teach collocations and MWIs are necessary for learners to acquire the required language items. All these help students become aware of gaps between their interlanguage and the language they are learning.

Explicit Instruction**:** There is a need for explicit instruction on a wider variety of lexical bundles. Teachers should focus on increasing learners' awareness of different types of bundles in different genres and their

appropriate contexts. Explicit instruction on features specific to writing as well as essay-organizing structures would be beneficial for writers at all levels. In this way, learners may better understand the patterns and usages and integrate those patterns into their own productions.

***Contextualized Practice*:** Providing learners with contextualized practice and exposure to authentic texts can help them understand and use a broader range of lexical bundles effectively. Learners can also compare their productions with that of native speakers, consult a learner corpus, and correct characteristic interlanguage issues such as underuse, overuse, misuse, all of which can enhance their writing skills. When learners refer to the concordance lines to see whether the structures they use are correct, concordances can reinforce the learning process.

Corpus-Based Learning: Using learner and native corpora in teaching can highlight gaps in learners' use of lexical bundles and provide models of native-like usage. Learners can also become aware of their errors and correct themselves thanks to various corpus tools provided. Data-Driven Learning (DDL) is a powerful approach that enhances language learning by providing learners with authentic language exposure, promoting self-correction, fostering discovery learning, and developing analytical skills (Gilquin & Granger, 2010). It transforms learners into active participants in their language acquisition process, making it a valuable method in language teaching. Several studies in the literature utilizing DDL in the teaching of lexical bundles in the classroom report positive findings.

The integration of the corpora applications into the English language teaching may be realized by using the freely available and user-friendly web-based tools that have become available such as Lancaster University BNClab (Gablasova, 2020), BNC (2014). Furthermore, some of the platforms are interactive such as FLAX (Interactive Language Learning: FLAX library (nzdl.org). As for the teachers who do not have any equipment and the Internet in their classrooms, some resources offer printed materials for DDL such as Tim Johns' Kibbitzers (https://lexically.net/TimJohns/). These software let students discover the language by themselves, see the words in their contexts, expose authentic language data, enabling them check their errors and use them while doing their homework. Teacher can make use of these tools in

the classrooms to see the number of occurrences of the words. They may also create their own corpus from their learners' writings or make their learners become aware of various text types and genres.

The results of these kind of studies should be integrated into ESL/EFL curriculums. Syllabus design can benefit from the use of specialized corpora in language teaching, tailored to students' proficiency levels and specific needs. Language teachers, ELT publishers and practitioners need to be enlighted about the utilization of corpus data and corpus-based teaching methods within their classrooms. Textbooks and other teaching materials should be prepared taking these frequently used structures into account. This can help students gain a deeper understanding of these structures and assist them in becoming more like native speakers. The incorporation of MWIs and bundles into the current syllabus should be promoted, and learners should be encouraged to verify their own language usage by referencing existing corpora. Last but not least, EFL teachers should be equipped with the skills of applying data-driven and corpus-based teaching techniques to use in their classrooms.

Future research should continue to explore the use of lexical bundles across different genres and proficiency levels and with learners from different L1 backgrounds. Additionally, investigating the impact of explicit instruction and corpus-based learning on learners' use of lexical bundles can provide valuable insights for language teaching.

Limitations of the Study

This study makes use of corpus-based techniques to analyze the use of lexical bundles in the written performances by Turkish EFL students and available in International Corpus of Learner English (ICLE) as a sub corpus. Although the corpus data to be used for this purpose is the most suitable one available in the field, it is still limited to the ones included in the TICLE and the LOCNESS. Also, only 3 and 4-word bundles were analysed in the study, excluding the erroneous structures. In addition, the study has been carried out by limiting its scope to the structural and functional categorization of lexical bundles produced by Turkish EFL learners and native speakers.

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*(2), 81–92. https://doi.org/10.1016/j.esp.2011.08.004

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software

Bal Gezegin, B. (2019). Lexical bundles in published research articles: A corpus-based study. *Journal of Language and Linguistic Studies, 15*(2), 520–534. https://doi.org/10.17263/jlls.586188

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

Biber, D., Conrad, S., & Cortes, V. (2003). Lexical bundles in speech and writing: An initial taxonomy. In A. Wilson, P. Rayson & T. McEnery (Eds.), *Corpus linguistics by the Lune: a festschrift for Geoffrey Leech* (pp. 71–93). Peter Lang. https://doi.org/10.1016/j.esp.2006.08.003

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., & Barbieri, F. (2007). Lexical Bundles in University spoken and written registers. *English for Specific Purposes, 26,* 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, D., & Conrad, S. (2009). *Real grammar. A corpus-based approach to English*. Pearson Longman.

Bychkovska, T., & Lee, J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes, 30*, 38–52. https://doi.org/10.1016/j.jeap.2017.10.008

Byrd, P., & Coxhead, A. (2010). On the other hand: Lexical bundles in academic writing and in the teaching of EAP. *University of Sydney Papers in TESOL,* 5, 31–64.

Chen, Y. 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

Chenu, F., & Jisa, H. (2009). Reviewing some similarities and differences in L1 and L2 lexical development. A*cquisition et interaction en langue étrangère,* 17–38. https://doi.org/10.4000/aile.4506

Conklin, K., & N. Schmitt. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72–89. https://doi.org/10.1093/applin/amm022

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

Cortes, V., & Lake, W. (2023). A solution to the problem of attaining observation independence in lexical bundle studies. *International Journal of Corpus Linguistics, 28*(2), 263–277. https://doi.org/10.1075/ijcl.21100.cor

Elturki, E. (2015). *The development of formulaic sequences: A longitudinal learner corpus investigation.* Dissertation, Washington State University, Washington. https://doi.org/10.1075/ijcl.18.1.07odo

Gablasova, D. (2020). *How corpus-based resources, BNClab and LancsBox, can be used by teachers to develop innovative classes that integrate digital technology.* Digital technology and innovation in teaching, Cork, Ireland.

Gilquin, G., & Granger, S. (2010). *How can data-driven learning be used in language teaching?* In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 359–370). Routledge.

Granger, S. (1998). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. P. Cowie (Eds.), *Phraseology: Theory, analysis and applications* (pp. 145–160). Clarendon Press. https://doi.org/10.1093/oso/9780198294252.003.007

Granger, S., Dagneaux, E., & Meunier, F. (2009). *The international corpus of learner English version 2. handbook and CD-ROM*. Louvain-la-Neuve: Presses Universitaires de Louvain. http://hdl.handle.net/2078.1/75579

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: a study into

the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics, 4*(2), 237–258. https://doi.org/10.1111/j.1473-4192.1994.tb00065.x

Hyland, K. (2008). 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

Karabacak, E., & Qin, J. (2013). Comparison of lexical bundles used by Turkish, Chinese, and American university students. *Procedia-Social and Behavioral Sciences*, *70*, 622–628. https://doi.org/10.1016/j.sbspro.2013.01.101

Morris, L., & Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System 32*, 75–87. https://doi.org/10.1016/j.system.2003.05.001

Muşlu, M. (2018). Use of stance lexical bundles by Turkish and Japanese EFL learners and native English speakers in academic writing. *Gaziantep University Journal of Social Sciences, 17*(4). https://doi.org/10.21547/jss.444386

Nekrasova, T. M. (2009). English L1 and L2 Speakers' Knowledge of Lexical Bundles. *Language Learning*, *59*(3), 647–686. https://doi.org/10.1111/j.1467-9922.2009.00520.x

Nesselhauf, N. (2004). 'Learner Corpora and their Potential for Language Teaching'. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 125–52). John Benjamins. https://doi.org/10.1075/scl.12.11nes

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

Ping, P. (2009). A study of the use of four-word lexical bundles in argumentative essays by Chinese English majors—a comparative study based on WECCL and LOCNESS. *CELEA Journal, 32*(3), 25–45.

Salazar, D. (2014). *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching* (Vol. 65). John Benjamins Publishing Company. https://doi.org/10.1075/scl.65

Schmitt, N., & Carter, R. (2004). Formulaic sequences in action: An introduction. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing, and use* (pp. 1–22). Benjamins.

https://doi.org/10.1075/lllt.9.02sch

Schmitt, N., Dörnyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (Ed.), *The acquisition and use of formulaic sequences* (pp. 55–86). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.9.05sch

Shin, Y. K. (2018). *Lexical bundles in argumentative essays by native and nonnative English-speaking novice academic writers.* Dissertation, Georgia State University. https://doi.org/10.57709/11994284

Sinclair, J. M. (Ed.). (1987). *Looking up: An account of the COBUILD project in lexical computing*. Collins ELT.

Staples, S., Egbert, J., Biber, D., & McClair, A. (2013). Formulaic sequences and EAP writing development: Lexical bundles in the TOEFL iBT writing section. *Journal of English for Academic Purposes, 12*, 214–225. https://doi.org/10.1016/j.jeap.2013.05.002

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* Benjamins. https://doi.org/10.1075/scl.6

Uçar, S., & Zarfsaz, E. (2022). A corpus-based study: The employment of lexical bundles in Turkish students' academic writing. *Türkiye Bilimsel Araştırmalar Dergisi, 7*(2), 329–341.

Wei, Y., & Lei, L. (2011). Lexical bundles in the academic writing of advanced Chinese EFL learners. *RELC Journal, 42*(2), 155–66. https://doi.org/10.1177/0033688211407295

Williamson, G. (2013). SLT info. http://www.sltinfo.com/type-token-ratio.html

Wood, D. (2015). *Fundamentals of formulaic language: An introduction.* Bloomsbury Publishing.

Sibel Aybek
Čukurovos universitetas, Turkija
sibelaybek1@gmail.com
Cem Can
Čukurovos universitetas, Turkija
cemcan01@gmail.com

LYGINAMOJI LEKSINIŲ JUNGINIŲ ANALIZĖ AKADEMINIUOSE RAŠTO DARBUOSE, KURIUOS PARENGĖ GIMTAKALBIAI ANGLŲ KALBOS VARTOTOJAI IR TURKAI, BESIMOKANTYS ANGLŲ KAIP UŽSIENIO KALBOS

Anotacija. Autentiška kalbos vartosena dažnai susideda iš pasikartojančių junginių, vadinamų daugiažodžiais vienetais, arba formulinėmis frazėmis (Byrd & Coxhead, 2010), kurios yra kaip esminiai „diskurso statybiniai blokai tiek žodiniame, tiek rašytiniame registre" (Biber & Barbieri, 2007, p. 263). Leksiniai junginiai, formulinių frazių sekos dalis, apibrėžiami kaip „pasikartojantys išsireiškimai, nepriklausomai nuo jų idiomatiškumo ir nuo jų struktūrinio statuso" (Biber ir kt., 1999, p. 990). Tyrimas nagrinėja dažniausiai vartojamų 3 ir 4 žodžių leksinių junginių vartojimą Tarptautinio besimokančiųjų anglų kalbos tekstyno (ICLE) turkų kalbos dalyje (TICLE) ir Louvain (Luvėno) gimtakalbių anglų kalbos rašinių tekstyne (LOCNESS) kaip kontroliniame lygiagrečiame tekstyne. Leksiniai junginiai klasifikuojami pagal jų struktūrines ir funkcines charakteristikas remiantis Biber et al. (2003; 2004) sukurta taksonomija. Atliekama interpretatyvi kontrastyvinė gimtakalbių (LOCNESS) ir negimtakalbių (TICLE) duomenų rinkinių analizė. Rezultatai rodo, kad anglų kaip užsienio kalbos besimokantys turkai vartoja per daug veiksmažodžių frazių fragmentų, o daiktavardžių ir prielinksninių frazių fragmentų – per mažai. Be to, TICLE tekstuose pastebima mažesnė leksinė įvairovė, palyginti su LOCNESS tekstais. Kalbant apie funkcinę klasifikaciją, anglų kalbos besimokantys turkai linkę per dažnai vartoti tam tikrą apibrėžtą funkcinių junginių dalį, nors apskritai jų vartoja mažiau. Šie rezultatai atskleidžia esmines problemas, kylančias mokant anglų kaip užsienio kalbos, ypač poreikį išsamiai ir tiksliai mokyti vartoti daugiažodžius vienetus.

Pagrindinės sąvokos: tekstynų lingvistika; besimokančiųjų tekstynai; leksiniai junginiai; daugiažodžiai vienetai; turkai, besimokantys anglų kaip užsienio kalbos.