Available online at www.jlls.org

JOURNAL OF LANGUAGE AND LINGUISTIC STUDIES

ISSN: 1305-578X

*Journal of Language and Linguistic Studies, 15*(2), 520-534; 2019

# Lexical bundles in published research articles: A corpus-based study

Betül Bal-Gezegin [a] *

[a] *Amasya University, Amasya, Turkey*

**APA Citation:**

Gezegin-Bal, B. (2019). Lexical bundles in published research articles: A corpus-based study. *Journal of Language and Linguistic Studies, 15*(2), 520-534.

Submission Date:31/10/2018

Acceptance Date:15/03/2019

**Abstract**

This corpus-based study investigates to what extent L1 Turkish speakers of English produce lexical bundles in their academic writing. To this end, a corpus of published research articles in six academic disciplines was collected. The corpus included one-million words in total. The four and five-word lexical bundles in the corpus were identified with the help of a corpus software and analyzed for their frequency, structural and functional features. The analysis yielded a total of 99 four-word and 22 five-word lexical bundles in the corpus. The results showed that the lexical bundles frequently used by Turkish authors in research articles had structural correlates and performed strong functions to construct the discourse of academic writing. Also, the study revealed a new group of bundles called research referential bundles. This finding might indicate that genre plays a significant role in the use of lexical bundles. The discussions given in this article could provide insights for further multi-word studies.



*Keywords:* Corpus; lexical bundles; research articles; academic journals, academic writing

## 1. Introduction

With the help of corpus tools, recent studies have paid attention to groups of words that occur together in both spoken and written languages (Nattinger & DeCarrico, 1992; Altenberg, 1998; Biber, Johansson, Leech, Conrad, & Finegan, 1999). Phraseology, which is explained as "the study of the structure, meaning and use of word combinations" (Cowie, 1994, p. 3168), has a rising status as a field of investigation. Idioms, collocations and lexical bundles are varieties of these combinations studied under the term phraseology. These multi-words have been referred to as lexical bundles/phrases (Nattinger & DeCarrico, 1992; Biber & Conrad, 1999; Biber et al., 1999; Stubbs, 2007a, 2007b), recurrent word combinations (Altenberg, 1998; DeCock, 1998), prefabricated patterns (Granger, 1998), clusters (Hyland, 2008a; Schmitt, Grandage & Adolphs, 2004), phrasal lexemes (Moon, 1998), and formulaic sequences (Schmitt & Carter, 2004), among others. The number of studies highlighting the significance

* Corresponding author. Tel.: 0358 211 50 05
*E-mail address*: betul.bal@amasya.edu.tr

** The information contained in this article was extracted from a master's thesis by the author, Betül Bal-Gezegin, for the Applied Linguistics and English as a Second Language Program at Georgia State University, Georgia, Atlanta.

of multi-words occurring frequently has been increasing particularly with the recent advances in computer assisted language investigations. The studies have revealed that LBs perform certain structural and discourse functions and these findings trigger further studies on formulaic language in different genres (i.e. spoken vs. written), languages, and contexts.

### *1.1. Literature Review*

Lexical bundles, also called n-grams, were first introduced by Biber et al. (1999) who described them as ‘‘recurrent expressions, regardless of their idiomaticity, and regardless of their structural status” (p. 990). Examples of lexical bundles in academic writing are expressions like “the end of the”, “as a result of” and “in the case of” (Biber et al. 1999). Lexical bundles differ from idioms and collocations in that they are “usually not complete structural units, and usually not fixed expressions” (Biber & Conrad, 1999, p. 183). In addition to being incomplete units (i.e. the presence of a), lexical bundles can combine several different parts of speech in a single string such as prepositions, nouns, or verbs. To be counted as a lexical bundle, a multiword expression needs to meet certain criteria about the frequency (i.e., occurring 20-40 times across five different texts in a million-word corpus, Biber et al., 2004; Cortes, 2004).

Recent studies have revealed that lexical bundles are found in many different registers both written and spoken. Academic prose is one the registers that has drawn attention in the use of lexical bundles (LBs, henceforth). Studies show that LBs have certain structural characteristics and various discourse functions in academic texts (Biber et al., 1999, 2003; Biber, Conrad & Cortes, 2004; Cortes, 2002, 2004). Previous studies mostly focused on defining the characteristics and functions of LBs in certain genres and compared LB use of English native speakers or other languages only (Cortes 2002, 2004; Scott & Tribble, 2006; Biber & Barbieri, 2007; Kim, 2009; Nesi & Basturkmen, 2006; Hyland, 2008a, 2008b). However, there has been an increase in the number of studies comparing LBs of native speakers (NS) and non-native speakers (NNS) (e.g. DeCock, 2000; Römer & Arbor, 2009; Chen & Baker, 2010; Ädel & Erman, 2012; Salazar, 2014). Chen and Baker (2010) analyzed lexical bundles found in NS and NNS students’ academic writing and found that there were variations between these two groups of texts in the use of lexical bundles. L2 students showed tendency to overuse certain lexical bundles such as all over the world and underuse expressions such as in the context of, which were found to occur frequently in academic prose. Similarly, Ädel and Erman (2012) compared the writings of NSs and NNSs and found a wider range of lexical bundle types in NSs’ writing. Based on the comparison of LB use in NS and NSSs’ written products, DeCock (2000) reported that in undergraduate writing, NNSs used more lexical bundles than NSs. Use of lexical bundles in expert and apprentice academic writings by NS and NNSs’ was also investigated by Römer and Arbor (2009) who reported that there were few differences between two groups, and both lacked academic English bundles. They concluded that learning the language requirements of academic writing (and phraseology) was a need for native speakers as well.

Previous studies found that LBs have certain functions as well as structural patterns and they can be divided into classes based on structure and their functions (Cortes, 2002; Hyland, 2008). Biber et al. (1999, p. 1015-1024) categorized the lexical bundles in academic genre in 12 major categories according to their structures. These are:

- Noun phrase with *of*-phrase fragment
- Noun phrase with other post-modifier fragment
- Prepositional phrase with embedded of-phrase fragment (*the end of the)*
- Other prepositional phrase fragment (*as in the case*)
- Anticipatory *it* + verb phrase/adjective phrase (*it is possible to*)
- Passive verb + prepositional phrase fragment (*is based on the*)
- Copula *be* + noun phrase/ adjective phrase (*is one of the*)

- (verb phrase +) that-clause fragment (*has been shown that*)
- (verb/adjective +) to-clause fragment (*are likely to be*)
- Adverbial clause fragment (*as shown in figure*)
- Pronoun/noun phrase + *be* (+…) (*there was no significant*)
- Other expressions (*as well as the*)

A framework for functions of LBs that has been commonly used in the related analysis was suggested by Cortes (2002, 2004) and further developed by Biber et al. (2004). In this taxonomy, LBs were divided into three functional groups; *stance expressions, discourse organizers* and *referential expressions*. The first group, stance bundles, are described as "overt expression of an author's or speaker's attitudes, feelings, judgments, or commitment concerning the message" (Biber et al., 2004, p. 386). Discourse organizers which help to structure the text have various sub-functions like presenting, clarifying and elaborating on the topic etc.

The last functional category, the referential expressions, which are not as common as the former two categories, relate to a specified attribute, a condition or refer to number, quantity, size as well as time and place. These three functional categories and their subcategories can be seen with examples in Table 1 below.

**Table 1.** Lexical Bundles' Functional Categorization (Biber et al., 2004, p.384)

| **Categories/Example** |
|---|
| **1. Stance Expressions** |
| A. Epistemic Stance |
| Personal: *I think it was* |
| Impersonal: *are more likely to* |
| B. Attitudinal/ Modality Stance |
| B.1) Desire: *if you want to* |
| B.2) Obligation/ Directive |
| Personal: *you look at the* |
| Impersonal: *it is necessary to* |
| B.3) Intention/Prediction |
| Personal: *what we are going to* |
| Impersonal: *is going to be* |
| B.4) Ability |
| Personal: *to be able to* |
| Impersonal: *it is possible to* |
| **2. Discourse Organizers** |
| A. Topic Introduction/Focus: *in this chapter we* |
| B. Topic Elaboration/Clarification: *on the other hand* |
| **3. Referential Expressions** |
| A. Identification/ Focus: *one of the most* |
| B. Imprecision: *and things like that* |
| C. Specification of Attributes |
| C.1) Quantity Specification: *a lot of people* |
| C.2) Tangible Framing Attribute: *in the form of* |
| C.3) Intangible Framing Attribute: *in the case of* |
| D. Time/Place/Text Reference |
| D.1) Place Reference: *in the United States* |
| D.2) Time Reference: *at the same time* |
| D.3) Text Deixis: *as shown in Figure N* |
| D.4) Multi-functional Reference: *at the end of* |

### *1.2. Overview of the present study*

There is no controversy that writers need to learn the conventions and requirements of academic genre which also includes the use of appropriate fixed lexical expressions (e.g., Cortes 2004, Biber, Gray & Poonpon 2013). Based on previous studies highlighting the importance of formulaic language in academic writing, this study focuses on LBs in published research papers written by Turkish academics in their L2, English. It aims to find the common four- and five- word lexical bundles in published research articles. The study is significant in that lexical bundle use has not been investigated in Turkish scholars' advanced writing (research articles) before and it also provides an answer to the question of how L2 English academic writers use of LBs in their writings produced for publishing internationally. The study is also significant in that there is no existing corpus with a size of one-million word which includes published English articles written by Turkish scholars. By investigating the use of lexical bundles as well as their structural and functional features based on earlier categories suggested by Biber et al. (1999) and Biber, Conrad and Cortes (2004), the study aims to extend the literature on phraseology. This study investigates the following research questions in order to achieve a thorough assessment of lexical bundles used by Turkish academics in English research papers:

1. What are the most prevalent four-and five- word lexical bundles found in Turkish scholars' published research articles?
2. What are the structural and functional characteristics of the lexical bundles identified?

## 2. Corpora and Method

There is no doubt that the use of computer tools in the study of lexical bundles have been crucial. Such tools help researchers to analyze and reach empirical conclusions based on their corpus collected. This study also used a computer software to analyze the texts gathered. The collection of data and the analysis based on taxonomies is explained in the following section.

### *2.1. Corpora*

As aforementioned, research articles from six different disciplines written by Turkish researchers have been collected from academic journals for the purpose of the study. The academic research article is one genre that has attracted considerable attention from researchers. Hyland (2012, p.159) describes the published research article as "the most discursively crafted and rhetorically machined genre" and shows it to be characterized by lexical bundles that function to "present research by engaging with a literature, providing warrants, establishing background, connecting ideas, directing readers around the text, and specifying limitations." For the purpose of this study, research articles written by Turkish authors and published in various academic journals were chosen as the target texts. It is thought that the outcomes of the research could be affected by including more than one sort of scholarly prose as lexical bundles are register-bound.

Regarding the size of the corpus for this research, the principle "A corpus must be large enough to adequately represent the occurrence of the features being studied" that is suggested by Biber (2006, p. 251) was followed. In this study, one-million-word corpus was collected from six different fields. Some disciplines did not have as many English articles written by Turkish scholars as others; therefore, availability of articles in the fields was the main criteria in the choice of six disciplines. As can be seen in the table 1 below, the corpus consisted of 200 articles in six different fields with a total number of 1,005,137 words.

**Table 2.** Turkish Scholars Research Articles Corpus

| Disciplines | # of words | # of articles |
|---|---|---|
| Economics | 164,745 | 29 |
| Education | 167,541 | 32 |
| History | 169,299 | 20 |
| Medicine | 153,715 | 44 |
| Psychology | 164,358 | 50 |
| Sociology | 185,479 | 25 |
| Total | 1,005,137 | 200 |

The texts gathered for the Turkish Scholars Research Articles Corpus (hereafter, TSRAC) were checked to ensure that the nationality of the authors was Turkish. Articles which had native speakers of English as co-authors were not included in the corpus collection. The journal selection process while creating the corpora was guided by experts (Turkish academics) who had already published in academic journals. They were asked to provide a list of journal names which were prestigious, international and were published in an English-speaking country.

### *2.2. Identification and Analysis of Lexical Bundles*

In order to consider a fixed group of word as a LB, it needs to meet certain predetermined criteria on frequency and distribution in a corpus. The threshold frequencies depend on the scope of each research and are arbitrary. In the literature, there are various ideas on the cut-off point of frequency ranging from 10 (Biber et al., 1999) to 40 times per million (Biber & Barbieri, 2007). In addition to its frequency, it also has to appear at least in five different texts in the corpus under investigation which ensures to prevent authors' idiosyncratic uses in the language samples. For the purpose of this study four- and five-word lexical bundles which appear at least 20 times across five different research articles in the corpus were retrieved. The reason of focusing on four-word bundles is because as Cortes (2004) put forward many three-word lexical bundles are already structurally a part of four-word bundles which are more common. Similarly, Hyland (2008) stated that compared to five-word bundles, four-word bundles were more common. They also show clearer structural and functional features compared to three-word bundles.

The retrieval of LBs in TSRAC based on the set criteria (a cut-off frequency of 20 and the cluster size of 4-word) was realized with the corpus tool called AntConc (Anthony, 2007). After the research articles were cleaned from non-textual and irrelevant information and sections such as tables, figures, graphs, numerical data (i.e. page numbers, formulations), and references, they were searched for four- and five-word LBs through *ngram* function of the corpus tool.

To define the bundles' structural and functional features, each bundle was analyzed qualitatively in its context via the concordance tool in AntConc. For the structural categorization, the structural taxonomy developed by Biber et al. (1999) described in section 1.2 was utilized. By splitting the constructions into two wider classifications as phrasal and clausal, a slight shift has been introduced to this model. Three subcategories have been differentiated for the phrasal bundles: "Noun-Phrase (NP) based," "Preposition Phrase (PP) based," and "Verb Phrase (VP) based." All clausal categories were grouped under the main category of "clausal". The revised version of the structural taxonomy can be seen in Table 3 below.

**Table 3.** Structural Taxonomy of Lexical Bundles

| Category | Example |
|---|---|
| **A. Phrasal** | |
| **1. NP-based** | |
| (connector +) NP with of- phrase fragment | *the end of the* |
| NP with other post modifier fragment | *the way in which* |
| **2. PP-based** | |
| PP with embedded of-phrase fragment | *as a result of* |
| Other Prepositional Phrase (fragment) | *at the same time, on the other hand* |
| **3. VP-based** | |
| Anticipatory it + VP/adjective P + comp. cl. | *it is possible to* |
| Passive verb +PPf | *is based on the* |
| Copula be + noun phrase/adjective phrase | *is one of the, is due to the* |
| Pronoun/NP + be | *this is not the, there are a number of* |
| **B. Clausal** | |
| (verb/adjective +) to-clause fragment | *is likely to be, to be able to* |
| (VP +) that-clause fragment | *should be noted that* |
| Adverbial clause fragment | *as shown in figure, if there is a* |
| **C. Other Expressions** | *as well as the* |

After structural analysis and categorization of lexical bundles, they were analyzed and classified for their functional features. For the functional categorization, the taxonomy developed by Cortes (2002) and further revised by Biber et al. (2004) was used. This taxonomy which includes three broad categories as stance expressions, discourse organizers and referential expressions can be seen in Table 1 with all its subcategories and examples for each.

## 3. Findings and Discussion

In this section, four-word and five-word lexical bundles that were identified in TSRAC are reported and their structural and functional features are discussed based on taxonomies. After the application of preset criteria about the identification of LBs, and the elimination of bundles that are not related (i.e. private names), the corpus tool yielded a list of 99 four-word and 22 five-word lexical bundles used by Turkish scholars in their published English research articles. There were five times more four-word LBs than five-word LBs. Since all the five-word bundles were longer sequences of the four-word bundles identified and four-word bundles had a higher frequency, the elaborate analysis and discussion is focused on four-word bundles in this article. All the five-word bundles and their frequencies in TSRAC can be seen in Table 4.

**Table 4.** Five-word Lexical Bundles in TSRAC

| Rank | Frequency | Range | N-gram |
|---|---|---|---|
| 1 | 56 | 35 | at the end of the |
| 2 | 54 | 42 | on the other hand the |
| 3 | 45 | 18 | according to the results of |
| 4 | 44 | 38 | one of the most important |
| 5 | 41 | 34 | of this study was to |
| 6 | 36 | 28 | as a result of the |
| 7 | 32 | 10 | the ministry of national education |
| 8 | 31 | 23 | the results of this study |

| | | | |
|---|---|---|---|
| 9 | 30 | 21 | at the beginning of the |
| 10 | 29 | 18 | on the basis of the |
| 11 | 29 | 25 | the aim of this study |
| 12 | 28 | 26 | the purpose of this study |
| 13 | 28 | 10 | the turkish version of the |
| 14 | 27 | 16 | in the case of the |
| 15 | 25 | 18 | to participate in the study |
| 16 | 22 | 18 | aim of this study was |
| 17 | 21 | 7 | it was determined that the |
| 18 | 21 | 10 | more than half of the |
| 19 | 21 | 14 | on the part of the |
| 20 | 20 | 16 | is one of the most |
| 21 | 20 | 19 | purpose of this study was |
| 22 | 20 | 5 | who participated in the study |

When the list of four-word LBs was retrieved through the computer tool (see Table 5 for the full list of four-word LBs in TSRAC), it was found that the most frequent four-word lexical bundles were on the other hand (151), the end of the (107), as well as the (88), in the case of (72) one of the most (67), all of which were also identified as frequent lexical bundles in the literature. According to Biber et al. (1999) the two most common four-word lexical bundles were in the case of (72) and on the other hand (151), which were also frequent in the TSRAC. 13 out 99 LBs (on the other hand, the end of the, as well as the, in the case of, one of the most, as a result of, at the end of, on the basis of, in terms of the, at the same time, was found to be, that fact that the, is one of the) appeared more than fifty times in one-million word, which showed a high-frequent use and the first 9 of these bundles 13 bundles had also been identified as frequent bundles by Biber et al. (2004), and Cortes (2004, 2008). When LBs identified in TSRAC were compared to the previously identified bundles, it was found that 48 of the total 99 lexical bundles had not been identified before Biber et al. (2004), and Cortes (2004, 2008). There were also LBs identified as highly frequent by the same scholars but not found in TSRAC such as in the absence of, the extent to which, in the presence of, and per cent of the). This might be due to the fact that although both TSRAC and the compared corpora were all from academic genre, the texts in the TSRAC were limited to research articles only. On the other hand, the corpus investigated by Biber et al. (2004) and Cortes (2004, 2008) included textbooks, student writings as types of academic texts.

In addition to identify the most frequent four- and five- word LBs in TSRAC, the other purpose of this study was to find the structural and functional features of the LBs. To do so, each bundle found was analyzed elaborately to find out its structural and functional features based on the taxonomies introduced earlier in this paper.

### *3.1. Structural Analysis of TSRAC Lexical Bundles*

The lexical bundles found in the TSRAC are not grammatically complete units as seen in the examples such as one of the most, the end of the, this study was to, to the results of etc. This finding agrees with the finding suggested by Biber et al. (1999) that in academic writing more than 95% of the lexical bundles were not complete units. The argument is further supported by Cortes (2004) 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).

The LBs found were categorized on their structural characteristics based on the taxonomy introduced earlier in Table 3. In this taxonomy, bundles are firstly grouped as phrasal or clausal. Phrasal bundles have subcategories of noun phrase-based (i.e. the end of the), prepositional phrase-based (i.e. at the end

of) and verb phrase-based (i.e. is one of the). Lexical bundles which incorporate dependent clauses are called clausal bundles (i.e. if we look at). The third group includes other expressions which was explained by Biber et al. (1999) as "lexical bundles that do not fit neatly into any of the other categories" (p.1024). Based on these categories, the LBs identified in the TSRAC are structurally categorized as seen in the Table 5 below.

**Table 5.** Structural Distribution of LBs in TSRAC

| PP (46) | NP (31) | VP (13) | Other Expressions (3) | CF Clause Fragment (2) |
|---|---|---|---|---|
| on the other hand | the end of the | was found to be | as well as the | that there is a |
| in the case of | one of the most | is one of the | as well as in | to be able to |
| as a result of | the fact that the | to participate in the | than half of the | |
| at the end of | a result of the | were found to be | | |
| on the basis of | the majority of the | are presented in Table | | |
| in terms of the | this study was to | it was determined that | | |
| at the same time | the rest of the | it was found that | | |
| in accordance with the | the results of the | it is necessary to | | |
| of this study was | the aim of this | are more likely to | | |
| in the present study | the role of the | it is possible to | | |
| according to the results | the beginning of the | were included in the | | |
| to the results of | the basis of the | participate in the study | | |
| of the most important | purpose of this study | that there was a | | |
| in the context of | the total number of | | | |
| at the time of | aim of this study | | | |
| on the part of | an important role in | | | |
| In addition to the | the case of the | | | |
| in the early #s | the purpose of this | | | |
| in line with the | results of the study | | | |
| in terms of their | the nature of the | | | |
| In this study the | the course of the | | | |
| of this study is | the #s and #s | | | |
| to the fact that | Turkish version of the | | | |
| of the fact that | an increase in the | | | |
| of the Ministry of | the second half of | | | |
| of the Turkish Republic | the size of the | | | |
| for the first time | the ways in which | | | |
| in the city of | the establishment of the | | | |
| in a study by | the characteristics of the | | | |
| in the form of | the relationship between the | | | |
| of the present study | the importance of the | | | |
| on the one hand | | | | |
| by the Ministry of | | | | |
| in the #s and | | | | |
| at the beginning of | | | | |
| in the late s | | | | |
| in the number of | | | | |
| with respect to the | | | | |
| at the level of | | | | |
| during the course of | | | | |
| in the face of | | | | |
| in the field of | | | | |
| of the patients were | | | | |
| at the University of | | | | |
| for the purpose of | | | | |
| with the help of | | | | |

The majority of the LBs found (46) were made up of PP and this is followed by 31 NP bundles. Examples of PP bundles are *in the context of, at the time of, on the part of* and examples of NP bundles are *the aim of this, the role of the, the beginning of the, the basis of the*. As seen in Table 5 the next category is VP and the bundles that were composed of verb phrases (13) were not as common as preposition and noun phrases. Some examples of VP from TSRAC are *was found to be, are presented in table, it is possible to* etc. Lexical bundles constructed with dependent clause fragments in TSRAC were only *that there is a,* and *to be able to*. Other bundles that did not fit in these categories were *as well as the, as well as in,* and *than half of the*. The overall distribution of LBs across structural categories can be seen in figure 1 below.

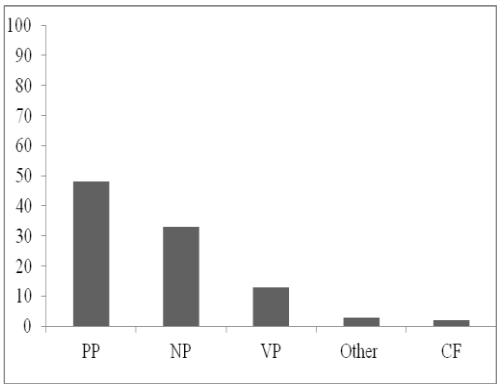


***Figure 1.*** Structural Distribution of TSRAC Lexical Bundles

The finding that LBs in academic research articles were primarily prepositional is in line with the results suggested by previous studies on LBs. Conrad and Biber (2004) reported that 60% of all lexical bundles they identified were structured as PP and NP in academic prose. It should be noted that verb phrase and dependent clause fragments, which were found to occur less in academic texts and more common in spoken language were also very few in academic research articles written by Turkish scholars.

### *3.2. Functional Analysis of TSRAC Lexical Bundles*

The LBs identified in TSRAC were analyzed and grouped into the categories of stance expressions, discourse organizers and referential expressions with their subcategories as seen in Table 1. The taxonomy used for the functions of LBs in TSRAC worked well for the categorizations; yet, slight changes were needed. For the referential expressions category, the subcategory called institute bundles proposed by Cortes (2008) was added. In addition, a new subcategory called research referential was created under referential expression category. Each category is further discussed with explanations and examples from TSRAC.

To begin with, stance bundles, which express personal feelings, attitudes, perspective, certainty, and uncertainty etc. (Biber, 2006), were less common (8%) compared to referential expressions and discourse organizers. There were two subcategories of stance bundles: epistemic and attitudinal. The former is about the certainty of the information and the latter shows the attitudes of the writer. Epistemic stance bundles found in TSRAC were the fact that the, to the fact that, of the fact that. Attitudinal stance bundles showing personal view points in the TSRAC were of the most important, to be able to, it is necessary to, are more likely to, it is possible to.

1) In fact, **it is possible to** argue that the paradox of Turkish nationalism enhanced the power of the state elites in Turkey and paved the way to a manufactured, official identity. (hist)

As the next main functional category of LBs, discourse organizers help to introduce, discuss and clarify a topic. It has two subcategories of topic introduction and topic elaboration. In TSRAC, %15 of the four word bundles functioned as discourse organizers and these bundles are on the other hand, as well as the, was/were found to be, in accordance with the, in the present study, as well as in, that there is/was a, in addition to the, it was determined that, it was found that, on the one hand, with respect to the, with the help of. Most of these bundles were used for elaboration and clarification purposes.

2) **On the one hand**, the absence of Cold War-era ideological polarization leaves space for tolerating the continuity of Soviet historiography; on the other, this continuity in method, conceptualization, and periodization is practical for historians trained in Soviet ideology and method. (hist.)

The third and the last group for the functional categorization of bundles is called referential expressions. Referential expressions are described as expressions referring to physical or abstract entities (Biber, 2004). It has sub-categories of identification/focus, imprecision, specification of attributes, and time/place/text reference. The present study revealed that most of the LBs found in TSRAC (75%) functioned as referential expressions. The taxonomy used for the categorization of LBs according to their functions needed two additional subcategories which are institution reference and research referential. The former has already been discussed and added to the taxonomy in a previous study by Cortes (2008). As its name suggests, the LBs in this group referred to an institution and examples from TSRAC for this subcategory are the Ministry of Education, of the Ministry of, by the Ministry of, at the university of etc. as can be seen in the example below.

3) At present, colleges of education train pre-school, elementary school, and secondary/high school teachers employed both by **the Ministry of Education** and the private schools, as well as inspectors for **the Ministry of Education**. (edu)

The latter subcategory added, namely research referential, is similar to what Hyland (2008a) suggested as his research-oriented category. The research referential bundles in the TSRAC were observed to give information about the research itself by focusing on the purpose, procedure, results, or participants of the study (i.e. aim of this study, to participate in the, were included in the). These bundles were different from text deixis in that text deixis referred to the paper (article or report) that presented the study and not to the study/research itself. However, research referential bundles in TSRAC behaved as meta-study expressions in that they directly addressed to the study conducted and referred to the actions needed to carry out the study as can be seen in the example 4.

4) **The aim of this study** is to examine the history of inflation accounting and its applications in Turkey. (econ)

The revised taxonomy used for identifying the LBs as referential expressions and the LBs found in TSRAC for each category can be seen in Table 6. In the overall analysis of lexical bundles, it was found that each type of referential bundles occurred in the TSRAC except for the category of imprecision.

**Table 6.** Lexical Bundles as Referential Expressions in TSRAC

| |
|---|
| A. Identification/ Focus*: one of the most, is one of the* |
| B. Imprecision |
| C. Specification of Attributes |
| C.1) Quantity Specification: *the majority of the, the rest of the, the total number of, for the first time, than half of the, the second half of* |
| C.2) Tangible Framing Attributes: *on the part of, in line with the, the size of the* |

C.3) Intangible Framing Attributes: *in the case of, as a result of, on the basis of, in terms of the, as a result of the, the beginning of the, in the context of, the basis of the, an important role in, the case of the, in the case of, in terms of their, the nature of the, the course of the, in the form of, an increase in the, Turkish version of the, the ways in which, in the number of, the establishment of the, at the level of, in the face of, in the field of, the characteristics of the, the relationship between the*

D. Time/Place/Text Reference

D.1) Place/event Reference: *in the Ottoman Empire, in the city of*

D.2) Time Reference: *at the same time, at the time of, in the early #s, the #s and #s, in the #s and, in the late #s, during the course of*

D.3) Text Deixis: *of this study was, this study was to, according to the result, are presented in Table, of the present study*

D.4) Institution*: the Ministry of Education, of the Ministry of, of the Turkish Republic, Ministry of National Education, by the Ministry of, at the university of*

D.5) Multi-functional Reference*: the end of the, at the end of, of the Ottoman Empire, at the beginning of*

E. Research Referential*: to participate in the, to the result of the, the results of the, the aim of this, purpose of this study, aim of this study, the purpose of this, results of this study, the results of the, in this study the, of this study is, in a study by, of the patients were, were included in the, for the purpose of, participate in the study*

---

When we look at the overall proportion of LBs across functional categories, it is seen that referential expressions constitute the largest part (%75) followed by discourse organizers (%15) and stance bundles (%8). When compared to other studies with a focus on LBs, it was found that these proportions vary. Chen and Baker (2016) reported that the majority of the LBs functioned as discourse organizers in their corpus of essays written by L1 Chinese learners of English. Similarly, Staples et al. (2013) revealed that in a corpus of EAP writing samples more than half of the LBs were in the category of discourse organizers. A recent study by Barbieri (2018) showed that in blogs the largest part of the LBs are stance bundles followed by referential expressions second, and discourse organizers as the least. Similar to the findings of the present study, Ädel and Erman (2012) also found that the largest part of proportion functioned as referential expressions in the academic writing samples they collected. These variations in the proportions of LBs' functions might indicate that different genres have different writing conventions. If it is academic writing, most of the bundles function as referential expressions. As seen in Barbieri's study (2018), for example, when writing is more personal as blogs, stance bundles that show writer's attitudes are more common. Pervasive stance bundle use was observed in a study by Herbel-Eisenmann, Wagner and Cortes, (2010) who anaylzed spoken corpus of mathematics classes with a focus of LBs. Since this corpus in this study was based on research articles as texts of academic writing, the finding of this study that a higher proportion of LBs functioned as referential expressions compared to stance bundles or discourse organizers supports the argument that genre plays a significant role in the use of lexical bundles and bundles also play an important role in constructing discourse.

## 4. Conclusion

By answering two main research questions, this study reached findings on LB use in academic texts written by Turkish scholars. Firstly, with the help of a corpus software, 99 four-word and 22 five-word lexical bundles were identified in a corpus of published research articles written by L1 Turkish speakers. All of the five-word LBs were already longer versions of 22 four-word bundles; therefore, the analysis was focused on four-word bundles which were more common. The top frequently used four-word bundles (appeared more than 50 times) were consistent with the bundles identified in previous studies on academic genre (Biber et al., 2004; Cortes, 2004, 2008). It was found that 53 of 99 four-word LBs

had not been identified in these previous studies on which the present study is based on. There might be several reasons for this case. The reason could be because the corpus used for this study is only research articles (excluding other types of academic texts such as textbooks). Another reason could be related to transfer from L1. It might be the case that these bundles unique to TSRAC could be English equivalents of lexical bundles used in L1, Turkish. Further investigation and follow up studies are needed to provide answers.

The next research question was about the structural and functional features of the LBs identified. To begin with, both the structural and functional taxonomies used from Biber et al. (1999, 2004) served the purpose of categorizing LBs in the present study, yet slight changes were needed. The largest group of bundles were structured with a preposition followed by nouns and verbs. In regard to the functions of LBs, a new sub-category called research referential bundles was created under the main category of referential bundles in the existing taxonomy. The finding that there was a new group of bundles used for referring to the study itself supports the argument that different genres have their own writing conventions. The texts in this research were all academic texts; therefore, coming up with a group of bundles related to research was a new but not an unusual result.

In further studies, the use of lexical bundles by novice and expert Turkish writers can be investigated. In addition, studies which aim to create awareness of L2 users in LB use in Turkish context is needed to see whether explicit teaching of fixed word combinations can help to use them in writing.

## Acknowledgements

I would like to thank to Dr. Viviana Cortes at Georgia State University who inspired me to investigate lexical bundles and provided insightful comments in each phase of this study.

### 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.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word combinations. In A. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 99–122). Oxford: OUP.

Biber, D. (1996). Investigating language use through corpus-based analyses of association patterns. *International journal of Corpus linguistics*, 1(2), 171–198.

Biber, D. (2006*). University language: A corpus-based study of spoken and written registers*. Amsterdam: Benjamin.

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

Biber, D., & Conrad, S. (1999). Lexical Bundles in Conversations and Academic Prose. In H. Hasselgard & S. Oksefjell (Eds.), *Out of corpora: studies in honour of Stig Johansson* (pp. 181–190). Amsterdam: Rodopi.

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). Frankfurt: Peter Lang.

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

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

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conver- sation to measure grammatical complexity in L2 writing development? TESOL Quarterly, 45, 5-35. doi: 10.5054/ tq.201 1.24448

Butler, C. (1997). Repeated word combinations in spoken and written text: Some implications for functional grammar. In C. Butler, J. Connolly, R. Gatward, & M. Wismans (Eds.), *A fund of Ideas: Recent development in functional grammar* (pp. 60–77). Amsterdam: Institute for Functional Research into Language and Language Use.

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

Cortes, V. (2002). Lexical bundles in Freshman composition. In R. Reppen, S. M. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 131–145). Amsterdam: John Benjamins Publishing Company.

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

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

Cowie A. P. (1994). Phraseology. In Asher, R.E. (ed.) *The Encyclopedia of Language and Linguistics*, 3168-3171.

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

DeCock, S. (2000). Repetitive phrasal chunkiness and advanced EFL speech and writing.In Mair, C., and Hundt, M. (Eds.), Corpus linguistics and linguistic theory (pp. 51-68), Amsterdam: Rodopi.

Erman, B. & B. Warren. (2000). The idiom principle and the open-choice principle, *Text*, 20, 29–62.

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

Güngör, F., & Uysal, H. H. (2016). A comparative analysis of lexical bundles used by native and non-native scholars. *English Language Teaching*, *9*(6), 176-188.

Granger, S., & Meunier, F. (Eds.). (2008). *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

Herbel&Eisenmann, B. Wagner, D. & Cortes, V. (2010). Lexical bundle analysis in mathematics classroom.discourse:The significance of stance Educational Studies in Mathematics, 75(1), 23-42.

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

Hyland, K. (2008b). Academic clusters: text patterning in published and postgraduate writing. *International Journal of Applied Linguistics*, 18, 41-62.

Hyland, K. (2012). *Disciplinary identities: Individuality and community in academic discourse*. Cambridge: Cambridge University Press.

Kim, Y. (2009). Korean lexical bundles in conversation and academic texts. *Corpora*, 4, 135-165.

Moon, R. (1998). *Fixed Expressions and idioms in English*. Oxford: Oxford University Press.

N. Schmitt (Ed.), Formulaic sequences acquisition, processing, and use (pp. 1- 22). Amsterdam; Philadelphia: John Benjamins Pub.

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

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signaling in academic lectures. *International Journal of Corpus Linguistics*, 11, 283-304.

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

Römer, U. & Arbor, A. (2009). English in academia: Does nativeness matter? Anglistik: International Journal of English Studies 20(2), (89-100).

Salazar, D. (2014). Lexical Bundles in Native and Nonnative Scientific Writing: Applying a Corpus-based Study to Language Teaching. Studies in corpus linguistics. John Benjamins Publishing Company, Amterdam/Philadelphia.

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

Schmitt, N., Grandage, S., & Adolphs, S. (2004). Are corpus-derived clusters psycholinguistically valid? In N. Schmitt (Ed.), *Formulaic sequences* (pp. 127–151). Amsterdam: Benjamins.

Scott, M., & Tribble, C. (Eds.). (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education*. Amsterdam and Philadelphia: John Benjamins B.V.

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

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

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

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

# Yayınlanmış araştırma makalelerinde sözcük öbeklerinin yapısal ve işlevsel analizi

**Öz**

Bu çalışma, Türk akademisyenleri tarafından İngilizce yazılmış araştırma makalelerinde bulunan sözcük gruplarının sıklık, yapı ve fonksiyonlarının derlem tabanlı incelemesidir. Araştırmanın amacı için, altı farklı akademik disiplinde yayınlanan bir milyon kelimelik araştırma makalesi derlemi oluşturulmuştur. Derlemdeki dört kelimeli sözcük öbekleri hem yapısal hem de işlevsel olarak tanımlanmış ve analiz edilmiştir. Araştırmanın bulguları, Türk yazarların araştırma makalelerinde sıklıkla kullandıkları sözlüklerin yapısal ilişkilere sahip olduğunu ve akademik yazım söylemini oluşturmak için güçlü işlevler gerçekleştirdiklerini ortaya koymuştur. Ayrıca, çalışma alanyazında daha önce tespit edilmemiş sözcük öbeklerini ortaya çıkarmıştır. Bu çalışmadan elde edilen sonuçlar, akademik türlerin ve sözcük gruplarının öğretilmesine uygulanabilir. Elde edilen bulgular ve tartışmalar, akademik yazıda daha fazla derlem temelli çalışmalara ışık tutabilir.

*Anahtar sözcükler*: Derlem; araştırma makaleleri, sözcük öbekleri; akademik yazım

**AUTHOR BIODATA**

Betül Bal Gezegin works as an Assistant Professor in the department of English Language Teaching at Amasya University, Turkey where she mainly teaches courses of research methods, language acquisition, material development, CALL and methodology courses. She is currently the head of Foreign Language Education and English Preparatory School at the same university. She holds a PhD in Foreign Language Education at Middle East Technical University. She obtained her MA degree at Georgia State University, Applied Linguistics program in the USA as a Fulbright scholar. Her academic interests mainly lie within the domains of Corpus Linguistics, CALL and ESP/EAP. Her latest publications include book chapters and research articles on corpus-based investigation of academic writing, metadiscourse markers in writing, teaching writing, and corpora and language teaching.