NEVŞEHİR HACI BEKTAŞ VELİ ÜNİVERSİTESİ
Ara, bul.

**Sosyal Bilimler Enstitüsü Dergisi**
**Cilt: 13 Sayı: 2 Haziran 2023**
**E-ISSN: 2149-3871**

# COMPARISON OF LEXICAL BUNDLES IN DISSERTATIONS CATEGORIZED BASED ON ACADEMIC DISCIPLINES

# AKADEMİK DİSİPLİNE GÖRE KATEGORİZE EDİLMİŞ DOKTORA TEZLERİNDEKİ SÖZ ÖBEKLERİNİN KARŞILAŞTIRILMASI

**Mustafa YILDIZ**
Samsun Üniversitesi
Yabancı Diller
myildiz55@yahoo.com
ORCID: 0000-0002-7971-5847

**Geliş Tarihi:**
31.01.2023

**Kabul Tarihi:**
14.06.2023

**Yayın Tarihi:**
30.06.2023

**Anahtar Kelimeler**
Söz öbeği, çok kelimeli ifadeler, kalıplaşmış ifadeler, derlem dilbilim

**Keywords**
Lexical bundles, multi-word expressions, formulaic expressions, corpus linguistics

## ABSTRACT

The present study aims to compare PhD dissertations, written by Turkish postgraduate students learning English as a foreign language, categorized based on the academic disciplines, in terms of the use of 4-word lexical bundles. To retrieve recurrent lexical bundles and to make their structural and functional analysis, various disciplines are grouped under two separate groups based on the academic fields such as hard and soft sciences. Also, English-major and non-English-major disciplines are compared to each other to see the variation in use of lexical bundles across disciplines. The findings reveal that ELT dissertations, representative of English-major disciplines, have three and four times as many lexical bundles as the dissertations written in the academic fields of soft and hard sciences, respectively. However, the academic fields produce almost the same number of lexical bundle types, suggesting that soft and hard sciences do not show variation in use of 4-word lexical bundles. With regard to the structural analysis of lexical bundles, it is found that lexical bundles most frequently appear in the syntactic categories of noun phrases and prepositional phrases. As for the functional distribution of lexical bundles, the findings indicate that the vast number of lexical bundles in each group function to be referential expressions.

## ÖZ

Bu çalışma, İngilizceyi yabancı dil olarak öğrenen Türk lisansüstü öğrencilerinin akademik disipline göre kategorize edilmiş doktora tezlerini 4 kelimelik sözcük demetlerinin kullanımı açısından karşılaştırmayı amaçlamaktadır. Tekrarlayan söz öbeklerini elde etmek ve yapısal ve işlevsel analizlerini yapmak için çeşitli disiplinler, fen bilimleri ve sosyal bilimler gibi bilimsel alanlara dayalı olarak iki ayrı grup altında toplanmıştır. Ayrıca, uzmanlığı İngilizce olan ve İngilizce olmayan disiplinler söz öbeklerinin kullanımındaki çeşitliliği görmek için birbirleriyle karşılaştırılmıştır. Bulgular, uzmanlığı İngilizce olan disiplinleri temsil eden İngiliz Dili Eğitimi doktora tezlerinin, sosyal ve fen bilimleri alanlarında yazılan tezlerden sırasıyla üç ve dört kat daha fazla söz öbeğine sahip olduğunu ortaya koymaktadır. Bununla birlikte, sosyal ve fen bilimleri alanlarında yazılan doktora tezleri hemen hemen aynı miktarda söz öbeği içermektedir, bu da sosyal ve fen bilimlerinin 4 kelimelik söz öbeklerinin kullanımında değişiklik göstermediğini göstermektedir. Söz öbeklerinin yapısal analizi, söz öbeklerinin en sık isim tümceleri ve edat tümceleri içerisinde yer aldıkları sonucunu ortaya çıkarmıştır. Söz öbeklerinin işlevsel dağılımıyla ilgili ise bulgular, her bir alt derlemdeki çok sayıda söz öbeğinin gönderge ifadeleri olduğunu göstermektedir.

***DOI:*** https://doi.org/10.30783/nevsosbilen.1245616
***Atıf/Cite as:*** Yıldız, M. (2023). Comparison of lexical bundles in dissertations categorized based on academic disciplines. *Nevşehir Hacı Bektaş Veli Üniversitesi SBE Dergisi, 13*(2), 936-953.

## Introduction

In addition to the four basic language skills, vocabulary is included as one of the sub-skills fundamental to gain the identity of acting as an interlocutor in a conversation. However, vocabulary is a complex issue in language learning, which requires, as Nation (1990) addresses, gaining mastery over many different facets of lexical knowledge, among which collocations and associations of words point to the fact that vocabulary is more than a one-word linguistic structure. Emphasizing the syntagmatic association between a stimulus word and the sequential words following it, Schmitt (2000) and Schmitt and Carter (2004) mention that words are organized in groups rather than individually and in the same way in the mental lexicon of native speakers of a language. Also, Byrd and Coxhead (2010) state that corpus linguistics helps to reveal both fixed and semi-fixed expressions that are frequently repetitive in the language and indispensable in terms of being fluent and sounding normal. Regardless of that they are called by a plethora of different terminologies in the literature (lexical bundles by Biber et al., 1999; formulaic sequences by Wray, 2002; recurrent word combinations by Altenberg, 1998; multi-word units and lexical chunks by Thornbury, 2002), formulaic expressions are often produced in discourse, regardless of register or genre. Erman and Warren (2000), for instance, found that formulaic expressions constitute slightly more than 50% of language forms in the data created from the corpora of both spoken and written forms. This number escalates even higher in Altenberg's (1998) study of recurrent word combinations in the London-Lund Corpus of Spoken English, in which more than 80% of the words in the corpus somehow form part of the larger lexical structure.

Emphasizing the prevalence of formulaic expressions in our everyday speech, Wray (2002) asserts that the studies compiled under three distinct categories in the existing literature interestingly include contradictory results. Wray (2002) points out that some findings in the literature center around the finding that native speakers frequently use formulaic expressions during communication. Granger (1998) even takes this finding a step further, arguing that non-native speakers of English produce fewer formulaic expressions than native speakers of English. Underlining that it is stated in the literature that first language and second language learners easily handle with formulaic expressions even at the beginning of the language acquisition process, Wray (2002) states that, in a group of other studies in the literature, the difficulty experienced by L2 learners with intermediate and advanced language proficiency in the use of formulaic expressions is the biggest obstacle to that they do not sound like a native speaker. Wray (2002) suggests that it is valuable to question the notion that formulaic expressions become more difficult for second language learners to deal with as their proficiency in a foreign language develops.

Among these formulaic expressions, lexical bundles, coined by Biber et al. (1999) for the first time and defined as "the most frequently recurring sequences of words" (Biber & Barbieri, 2007, p. 264), are three or more lexical items which are used in combination with one another repeatedly. On the other hand, for example, is formed by combining four words and these four individual words frequently come together in many different text types to introduce contrasting ideas. Lexical bundles have some distinguishing features which set them apart from other formulaic expressions. According to Biber and Barbieri (2007) lexical bundles are too numerous in language use, semantically transparent and do not have to be structurally complete. Namely, lexical bundles are a type of formulaic expressions which are great in number and their meanings are a combination of the meanings of the individual words that compose them. Also, a group of words that do not have a complete syntactic formation such as phrases, clauses, or sentences can also be called as lexical bundles.

Lexical bundles have long been a question of great interest in a wide range of studies. Previous studies on lexical bundles have dealt with these formulaic expressions from many different aspects. The use of lexical bundles in many different contexts has been the subject of research in these studies. For example, in order to determine the frequencies of lexical bundles, both mega corpora and specialized corpora have been investigated. Research articles (Bal-Gezegin, 2019; Cortes, 2004; Cortes, 2008), PhD dissertations (Bao & Liu, 2022; Hyland, 2008), master's theses (Hyland, 2008), classroom teaching discourse (Biber et al., 2004), textbooks (Biber et al., 2004; Hsu, 2014; Üstünbaş & Ortaçtepe, 2016), and oral proficiency exams (Üstünbaş & Ortaçtepe, 2016) appear in lexical bundle studies as specialized corpus data. On the other hand, although the frequency of the lexical bundles by itself means a lot, they do not contain pedagogical implications and do not contribute much to the language teaching studies carried out in the classroom environment. Therefore, there are a number of studies

which investigate lexical bundles' structural and functional characteristics (Bal-Gezegin, 2019; Biber et al., 2004; Cortes, 2004, 2008; Hyland, 2008). The relevant literature also consists of studies comparing the use of lexical bundles by non-native speakers of English to those by native speakers of English (Allen, 2010; Ortaçtepe, 2013, Öztürk & Durmuşoğlu-Köse, 2016).

The present study will try to expand on the findings of the aforementioned studies by comparing PhD dissertations, written by non-native speakers of English, categorized based on the academic disciplines. In order to identify recurrent lexical bundles in PhD dissertations and to make their structural and functional analysis, various disciplines are divided into two separate groups based on the academic fields such as hard and soft sciences. Also, English-major and non-English-major disciplines will be compared to each other to see any potential difference between majors.

## Literature Review

The literature includes a large number of studies conducted on formulaic expressions which, for instance, show a correlation observed between students' fluency, overall proficiency scores, and their utilization of formulaic expressions (Üstünbaş & Ortaçtepe, 2016); dealing with the syntactic and functional analysis of formulaic expressions in learners' written outputs (Allen, 2010); bringing together the most frequent collocations of spoken English (Shin & Nation, 2007); investigating semantically opaque formulaic expressions in textbooks (Hsu, 2014); examining structural and functional features of lexical bundles (Bal-Gezegin, 2019; Güngör & Uysal, 2016; Hyland, 2008); comparing the lexical bundle use of academics with those of learners (Cortes, 2004); and analyzing the use of lexical bundles in the articles of a specific academic discipline in two different languages (Cortes, 2008). Üstünbaş and Ortaçtepe (2016) examined students' use of formulaic expressions in oral proficiency exam. Examining the effects of different task types on the use of formulaic expressions, Üstünbaş and Ortaçtepe (2016) also analyzed the relationship between students' use of formulaic expressions and fluency and overall scores in oral proficiency exam. The researchers made content analyses of the textbook the students used throughout the semester and of the spoken discourse of students' oral examinations in terms of the use of formulaic expressions based on Kecskes' formulaic continuum (2007). The results showed that the students used almost 60% of the formulaic expressions available in the textbook in the oral proficiency exam. The findings on whether the use of formulaic expressions differed according to the type of task indicated that students used significantly more formulaic expressions in paired tasks compared to individual tasks. As for the relationship between the use of formulaic expressions and students' fluency and overall scores in oral proficiency exam, the findings revealed that a significant correlation occurred between the use of formulaic expressions and students' fluency and overall proficiency scores, implying that as the use of formulaic expressions increases, students' speech flow becomes more fluent and their level of proficiency increases.

Allen (2010) investigated the use of formulaic language in learners' written outputs. Examining the type and frequency of formulaic expressions produced by Japanese EFL learners in ALESS Learner Corpus, Allen (2010) compared the learners' use of formulaic expressions with those who are the authors of published research articles or writers as native speakers of English by using three reference corpora. In learners' written outputs, 4-word formulaic expressions that have conjoined together at least 40 times were taken into account for the analysis. A total of 144 formulaic expressions were analyzed in terms of their accuracy, grammatical structures, and characteristic functions in the text. On the one hand, as for accuracy, the learners' written products contained the use of formulaic language with high level of accuracy. On the other hand, as for grammatical structure of formulaic expressions, NP+of appeared as the most frequently used syntactic category. Last, with regard to the functions of formulaic expressions, more than 60% of the formulaic expressions in the learners' outputs were research-oriented bundles which help the authors convey information about the investigation to the audience.

Shin and Nation (2007) created a useful list of the most frequent collocations of spoken English for elementary learners of English based on the data from ten-million-word BNC spoken section. The researchers used a number of criteria to determine these collocations. For example, each pivot word was treated as a word type, not a member of a word family because different word types available within a word family might collocate with various words. Also, in order to find meaningful sequences instead of the sequences of words with grammatical functions, Shin and Nation (2007) determined content words as pivot words to be searched for. Furthermore,

the pivot words to be searched for must be among the most frequent 1,000 content words in spoken English as the researchers tried to arrive at a collocational list for elementary learners of English, and less frequent pivot words might create the risk of an additional learning burden. With regard to the frequency of potential collocational sequences, Shin and Nation (2007) chose the word sequences which occur 30 times minimum in ten million words. Additionally, the collocations determined must be smaller meaningful immediate constituents within larger immediate constituents. Namely, for the sake of frequency of collocational sequences, the meaning in the word groups should not be overlooked. Lastly, a collocation might have a number of semantic senses. Therefore, each of those collocational units with various meanings must be counted separately. The findings revealed that the most frequent 1,000 words in spoken English formed 5,894 collocations in total in BNC spoken section. The more frequent the pivot word is, the greater the number of collocations it creates. For example, the number of collocations formed by the most frequent 100 pivot words was more than one third of the total number of collocations formed by the total of 1,000 pivot words. Also, two-word collocations were more than three-quarters of the total number of collocations.

Hsu (2014) made an attempt to elicit semantically opaque formulaic sequences in college textbooks used by non-English major academic disciplines. The researcher built a corpus of 200 textbooks comprised of many words slightly above 20 millions by using the textbooks used by a total of 40 non-English major academic disciplines. 2-, 3-, 4-, and 5-word sequences which occur 5 times per million words were considered as target formulaic sequences. Some further criteria were also determined to consider those sequences as target formulaic sequences: On the one hand, those aforementioned word sequences must appear in each of the 40 academic disciplines. On the other hand, any of those formulaic sequences must appear in at least half out of 200 textbooks. The findings indicated that a total of 475 semantically opaque formulaic expressions appeared among the data analyzed. Also, these 475 non-compositional formulaic sequences were formed of 1,248 individual tokens, almost 89% of which were among the most frequent first 1,000 word lists in the BNC/COCA word-frequency scale, indicating that even if they are non-compositional formulaic sequences, they are mostly formed of the most frequent words in English deceptively.

Ortaçtepe (2013) investigated how nativelike the formulaic expressions used by Turkish students who are studying as international students in the USA. It was tried to measure the development of students in terms of the use of formulaic expressions in a one-year period by using Discourse Completion Test in pre-test and post-test design. The use of formulaic expressions by non-native students was also compared with the use of formulaic expressions by native speakers of English. How native-like the formulaic expressions of non-native speakers was also evaluated by a group of raters. The results revealed that native speakers of English used more formulaic expressions. In addition, the formulaic expressions they used were considered more native-like by raters. Although they were not considered as native-like as native speakers, formulaic expressions used by non-native speakers in pre-test and post-test differed in terms of nativelikeness rating. Their nativelikeness scores in post-test significantly escalated.

Bal-Gezegin (2019) used a frequency-driven approach to determine the most frequent 4-word and 5-word lexical bundles in the research articles authored by Turkish scholars. The structural and functional characteristics of these lexical bundles were further examined. A total of 200 articles published in academic journals from 6 different disciplines constituted the database of the study with a total word count of slightly above one million. Word sequences appeared 20 times per million words in at least 5 different articles were retrieved. The results showed that there was a total of 121 lexical bundle types in the dataset, where 4-word sequences were 5 times the number of 5-word sequences. As for the structural analysis of these lexical bundles, it seems that they appeared mostly within prepositional phrases and noun phrases. Furthermore, functional analysis of these lexical bundles revealed that while an overwhelming number of these word sequences (75%) function as referential expressions, a very limited number of them (8%) function as stance bundles.

Hyland (2008) investigated 4-word lexical bundles and their structural and functional features in 4 different academic disciplines in a corpus formed of three and a half million words. Research articles, PhD dissertations and master's theses constituted the data of the research. Four-word lexical bundles, which appear 20 times per million and are available in 10% of the total number of texts examined, were retrieved. These lexical bundles attained in accordance with those predetermined criteria were also grouped structurally and functionally. The

results showed that in the corpus of three and a half million words, 240 types of 4-word lexical bundles repeated themselves approximately 16,000 times. Electrical engineering was the discipline in the first ranking in terms of the use of lexical bundles among 4 different disciplines, in which 213 types covered 3.5% of the total number of words. With regard to the structure of these lexical bundles, noun phrase with of-phrase fragment was the most typical grammatical structure in the overall data. As for the functions of the lexical bundles, text-oriented lexical bundles, which help organize the text, outnumbered the research-oriented and text-oriented lexical bundles.

Cortes (2004) compared the lexical bundle use of academics who have published in the fields of history and biology with the lexical bundle use of students studying in these two fields. After the 4-word lexical bundles in the published articles of the academics were determined, they were grouped structurally and functionally. These lexical bundles, which were identified in the published articles, were called target bundles as they would subsequently be investigated in the written texts of the students. The findings revealed that the students either rarely or never used the lexical bundles frequently used by published authors. Furthermore, some of the lexical bundles used by the students did not convey the same function even though they were the same lexical bundles used by published authors.

Biber et al. (2004) adopted a frequency-driven approach to identify lexical bundles available in classroom teaching discourse and textbooks. They further compared the findings to those lexical bundles found in conversation and academic prose, attained in previous research. The cut-off point for the inclusion of lexical bundles in the analysis was accepted as occurring at least 40 times per million. The lexical bundles identified were also analyzed structurally and functionally. The findings indicated that classroom teaching discourse contained the largest number of lexical bundles, even larger in number than those in conversation. Classroom teaching not only outnumbered the other registers in terms of the use of lexical bundles but also was clearly ahead of other registers in the use of each of the three different functional categories. The findings also revealed that the structure of a lexical bundle and the function it performed significantly overlapped.

Cortes (2008) investigated the use of lexical bundles in academic history writing. She made a comparative analysis on two different corpora, each of which contained history articles either written in English or Spanish. Identified 4-word lexical bundles which appeared 20 times per million words and at least in 5 different texts, these bundles were classified both structurally and functionally. The findings obtained in each corpus were also compared to each other to reveal to what extent history writing in different languages resembled or differentiated. The findings revealed that the lexical bundles available in Spanish history writing dramatically outnumbered those in English history writing. As for the inherent structure of the lexical bundles in each corpus, prepositional phrases and noun phrases formed the structure of these lexical bundles. The functional classification of these bundles indicated that these word sequences in both corpora mostly performed referential functions. Also, slightly more than one-fifth of the lexical bundles overlapped in both corpora. Even some of the verbs used before the lexical bundles were also the same, indicating that although they are written in different languages, there are some commonalities in history writing between English and Spanish in terms of the use of lexical bundles.

Güngör and Uysal (2016) compared the use of lexical bundles in research articles written by native and non-native speakers of English. They also investigated whether the use of LBs by non-native speakers of English differentiated from those by native speakers of English functionally and structurally. They further investigated the LBs both common to two groups and peculiar to native speakers of English. Two specialized corpora were formed of research articles in the field of educational sciences. The findings revealed that the number of LBs used by non-native speakers of English was slightly more than 3 times that of native speakers of English. While native speakers of English primarily used more phrasal lexical bundles, non-native speakers of English used LBs in clausal structures. In terms of functional sub-categories of LBs, native speakers of English mostly used research-oriented bundles while non-native speakers of English mostly used text-oriented LBs.

Bao and Liu (2022) searched for the lexical bundles which statistically differentiated across two groups of linguistics dissertation abstracts written by native and non-native speakers of English. 700 dissertation abstracts for each of the databases were retrieved. 3-word lexical bundles which occurred 60 times per million words at least 2% of the total number of texts (14 texts in each database) were retrieved. The findings revealed that

Chinese authors substantially used greater number of LBs than American authors. These two groups of authors also differentiated in terms of the use of LBs which showed substantial variation across two corpora. While 63,14% of the LBs frequent in Chinese authors' corpus were infrequent in American authors' corpus, 48,21% of the LBs frequently used in American authors' corpus were infrequent in Chinese authors' corpus. In terms of the functional distribution of LBs used differently, the results indicated that text-oriented bundles were the most frequent bundles but the distribution of bundles-used-differently was inequivalent across each of the functional categories. In terms of the rhetorical move, bundles-used-differently most frequently occurred within the move of result.

The studies mentioned above indicate that the existing literature on lexical bundles is extensive and focuses particularly on their use and functional and structural analysis either in different academic disciplines or in different genres. In addition, the studies above show that there is a tendency among researchers to compare the use of lexical bundles between native and non-native speakers of English. The present study will try to expand on the aforementioned studies by investigating non-native English authors' use of lexical bundles in PhD dissertations categorized based on both by the field of science and by the majors of authors. In order to retrieve recurrent lexical bundles available in PhD dissertations and to make further structural and functional analysis, a group of academic disciplines are categorized into two groups based on the academic fields such as hard and soft sciences. In addition, with the inclusion of ELT dissertations as a third group, English-major and non-English-major disciplines are compared to each other to see the potential differences in use of lexical bundles, if available. The following research questions will guide the current investigation:

1. Are there any significant differences among the sub-corpora, i.e., soft sciences, hard sciences, and English Language Teaching dissertations, in terms of the use of lexical bundles?
2. What are the linguistic structures of lexical bundles available in each sub-corpus?
3. What are the functional characteristics of lexical bundles in each sub-corpus?

## Method

The current study employs a small-scaled specialized corpus of three sub-corpora to answer the research questions. In order to identify recurrent lexical bundles in PhD dissertations written in English by non-native speakers of English, three different sub-corpora were created by means of randomly selected PhD dissertations, each of which was completed between the years 2021 and 2022. First of all, various disciplines were divided into two separate groups based on the academic fields such as hard and soft sciences. Namely, two of the three sub-corpora in the study were composed of PhD dissertations written in different disciplines as much as possible, separated as hard and soft sciences. While the soft science sub-corpus consisted of the sum of 922,005 word tokens with the inclusion of 16 dissertations, the hard science sub-corpus was formed of 16 dissertations with 723,082 word tokens. Table 1 shows the academic disciplines which provide dissertations as data to the current study.

**Table 1.** Distribution of dissertations across academic disciplines

| Soft Sciences | Hard Sciences |
|---|---|
| Tourism Management | Physics Engineering |
| History | Biotechnology |
| Archaeology and History of Art | Maritime Transportation Engineering |
| Demography | Electrical and Electronics Engineering |
| Anthropology | Physics |
| Philosophy | Civil Engineering |
| Economics | Statistics |
| Musicology | Chemistry |
| Organization Studies | Cryptography |
| Western Languages and Literature | Molecular Biology and Genetics |
| Political Science and Public Administration | Polymer Science and Technology |
| Sociology | Software Engineering |
| Psychology | Aeronautics and Astronautics Engineering |
| Translation and Interpreting | Agricultural Machinery and Technologies Engineering |

| | |
|---|---|
| International Affairs | Biology |
| Business Administration | Mathematics |

The third sub-corpus was composed of 16 PhD dissertations with a total of 887,920 word tokens written in the field of English Language Teaching. Thus, it has been tried to obtain findings about the recurrent use of lexical bundles both in different academic fields and between English-major and non-English-major disciplines.

## Procedure

There are different trends toward which recurrent lexical bundles should be retrieved depending on the number of words they contain. Although there are previous studies among others which have focused on 3-word (Bao & Liu, 2022; Dahunsi & Ewata, 2022) and 5-word lexical bundles (Bal-Gezegin, 2019 searching for both 4-word and 5-word lexical bundles), the general trend is to search for 4-word lexical bundles (Allen, 2010; Cortes, 2004; 2008; Hyland, 2008; Perez-Llantada, 2014). Cortes (2004) underlines that 4-word lexical bundles provide more fruitful data to analyze in terms of structural and functional analysis compared to 3-word and 5-word lexical bundles. In a similar vein, Chen and Baker (2010) define 4-word lexical bundles as "the most researched length for writing studies, probably because the number of 4-word bundles is often within a manageable size (around 100) for manual categorization and concordance checks" (p. 32). Also, Hyland (2012) underlines that "(f)our-word bundles seem to be most often studied, perhaps because they are over 10 times more frequent than five-word sequences and offer a wider variety of structures and functions to analyze" (p. 151). Accordingly, in the current research, recurrent 4-word lexical bundles are retrieved in the specialized corpus by means of AntConc Software (Anthony, 2022).

Another variable which previous studies differ is about which lexical bundles will be considered as data based on their frequency. In order to identify recurrent lexical bundles, past studies in the literature have determined many different cut-off points showing dispersion among 5 occurences (Hsu, 2014), 10 occurrences (Biber et al., 1999), 20 occurrences (Bal-Gezegin, 2019; Cortes, 2008; Hyland, 2008; Perez-Llantada, 2014), 25 occurrences (Öztürk & Durmuşoğlu-Köse, 2016), 30 occurrences (Shin & Nation, 2007), 40 occurrences (Allen, 2010; Biber et al., 2004), 50 occurrences (Dahunsi & Ewata, 2022) and 60 occurences (Bao & Liu, 2022) per million words. Biber and Barbieri (2007) state that "(t)he frequency cut-off used to identify lexical bundles is somewhat arbitrary" (p. 267). How big or small the corpus or database to be analyzed seems to have an effect on which lexical bundles will be retrieved as target data. Because it does not contain a large amount of data, the current study identifies 40 occurrences of lexical bundles per million words in each of the sub-corpora as a frequency criterion for retrieving lexical bundles.

Furthermore, in order to prevent to retrieve individual idiosyncratic use of lexical bundles, past studies have determined different criteria which generally range from the occurrence of lexical bundles in 5 different texts (Bal-Gezegin, 2019; Biber et al., 1999; Cortes, 2008) to at least in 10% of the texts (Hyland, 2008; Perez-Llantada, 2014), even to at least in 2% of the texts (Bao & Liu, 2022). In the current study, a distribution criterion, which requires lexical bundles to appear in at least 4 different texts, corresponding to at least 20% of the texts in each sub-corpus, was used in order to avoid the retrieval of individual idiosyncratic use of lexical bundles as much as possible.

Overlapping 4-word lexical bundles, contrary to Chen and Baker (2010) concerning reducing frequency inflation, are not merged as 5-word lexical bundles and considered separately as in Perez-Llantada (2014) because merging these bundles as 5-word lexical bundles and excluding them from the findings will cause to loss of some important data. For example, to merge 'on the other hand' and 'the other hand the' as 5-word lexical bundle to be 'on the other hand the' and to exclude the former two from the findings will cause the loss of 'on the other hand' which is the most frequent lexical bundle in each sub-corpus. Similarly, combining 'as a result of' and 'a result of the' into 'as a result of the' to be a 5-word lexical bundle and excluding the former two 4-word lexical bundles would result in the loss of 'as a result of', one of the most frequently used lexical bundles in each sub-corpus. However, the following content/context-dependent lexical bundles are excluded: a statistically significant difference, statistically significant difference between, significant difference between the, participants stated that they, the participants stated that in ELT sub-corpus; in the …th century in Soft Science sub-corpus.

The taxonomy for the structural categorization of lexical bundles in academic prose (Biber et al., 1999) is used to analyze the grammatical structures of the 4-word lexical bundles obtained as a result of the data analysis. The structural categories to be used to group the lexical bundles include noun phrase with of-phrase fragment, noun phrase with other post-modifier fragments, prepositional phrase with embedded of-phrase fragment, other prepositional phrase (fragment), anticipatory it + verb phrase/adjective phrase, passive verb + prepositional phrase fragment, copula be + noun phrase/adjective phrase, (verb phrase+) that-clause fragment, (verb/adjective+) to-clause fragment), adverbial clause fragment, pronoun/noun phrase + be + (+…), other expressions.

The examination of the discourse functions of 4-word lexical bundles constitutes the final analysis of the current study. The lexical bundles attained at the end of the analysis of each of the sub-corpora are subjected to the functional classification of lexical bundles, suggested by Biber et al. (2004), grouping the bundles into three: stance expressions, discourse organizers, and referential expressions. According to Biber et al. (2004), on the one hand, stance expressions encode the speakers' or writers' attitudes toward the issues or their authorial stance toward the certainty of the proposition. On the other hand, while discourse organizers help speakers/writers link and organize the previous and the remainder content, referential expressions are used in the text to refer to all of concrete or abstract elements or to one or some of them.

In order to see any significant differences are available among sub-corpora in terms of the use of lexical bundles, log-likelihood values are calculated. The critical LL value 15.13 and above are adopted to consider the difference among sub-corpora to be significant as Rayson et al. (2004) advise on the use of the highest cut-off Log-likelihood value (15.13), which implies that the findings are significant at 99.9% accuracy, as a result of the comparison of Chi-squared and Log-likelihood tests by emphasizing that "in order to extend applicability of the frequency comparisons to expected values of 1 or more, use of the log-likelihood statistic is preferred over the chi-squared statistic, at the 0.01% level. The trade-off for corpus linguists is that the new critical value is 15.13" (p. 926).

## Results

### Frequency Analysis of Lexical Bundles

The findings towards the frequency of lexical bundles across three sub-corpora are summarized in Table 2. The analysis of 4-word bundles in each sub-corpora reveals that the type frequency of 4-word lexical bundles in ELT dissertations is almost 4 times the number of lexical bundles in dissertations in hard sciences (78 vs. 22), and almost 3 times those in soft sciences (78 vs. 28).

**Table 2.** Type/Token frequency and corpus size in each corpus

| Sub-Corpora | Type Frequency | Token Frequency | Corpus Size |
|---|---|---|---|
| ELT | 78 | 6,050 | 887,920 |
| Hard | 22 | 1,584 | 723,082 |
| Soft | 28 | 2,132 | 922,005 |

Additionally, as shown in Table 2, while the token frequency of lexical bundles in ELT dissertations is 6,050, this number is 2,132 in dissertations written in soft sciences, and 1,584 in dissertations in hard sciences. In other words, while the ratio of 4-word lexical bundles to the total number of words in ELT dissertations is 0.68 occurrences per 100 words, it is 0.23 occurrences per 100 words in soft sciences and 0.22 occurrences per 100 words in hard sciences. Frequency analysis of the use of lexical bundles among sub-corpora by means of Log-likelihood calculator indicates that the overall use of 4-word lexical bundles in ELT dissertations is more frequent than both those in soft sciences and hard sciences with Log-Likelihood values +2105,87 and +1950,10, respectively ($p<.0001$), meaning that there is a statistically significant difference between ELT dissertations and those in soft sciences and hard sciences (critical LL value: 15,13). Namely, more frequent use of lexical bundles in ELT dissertations as compared with those in soft sciences and hard sciences is justified by LL calculations. Furthermore, although the ratio of the number of words of the 4-word lexical bundles in soft sciences to the total number of words in the sub-corpus is slightly higher than that of the hard sciences (0,23 vs. 0,22), any significant differences (2,66 LL value, $p > ,0001$) were found between the soft sciences and hard sciences in terms of the use of lexical bundles.

As a result of the analysis of 3 sub-corpora, 94 different lexical bundles were found. Table 3 shows the lexical bundles, common to all sub-corpora, and their frequency values.

**Table 3.** Frequency of lexical bundles common to all three sub-corpora

| Lexical Bundles | Hard Science | Soft Science | ELT |
|---|---|---|---|
| On the other hand | 235 | 396 | 336 |
| As a result of | 112 | 195 | 153 |
| Is one of the | 99 | 91 | 92 |
| One of the most | 78 | 90 | 92 |
| In terms of the | 68 | 45 | 87 |
| At the end of | 63 | 76 | 247 |
| The other hand the | 58 | 59 | 61 |
| The end of the | 52 | 111 | 212 |
| A result of the | 51 | 92 | 66 |
| At the same time | 50 | 88 | 80 |
| The results of the | 46 | 55 | 198 |

On the other hand, while some of the above-mentioned 94 lexical bundles are available in two of the three sub-corpora, they are absent in one. As Table 4 indicates, 10 of the 94 lexical bundles exist in the dissertations written in the field of soft sciences and ELT; however, they are not included in those in the field of hard sciences. While there is only one lexical bundle peculiar to the hard sciences and ELT sub-corpora but not included in the soft sciences sub-corpus, similarly, there is only one lexical bundle common in hard and soft sciences and is not included in the ELT sub-corpus.

**Table 4.** Frequency of lexical bundles just peculiar to two sub-corpora

| Lexical Bundles | Soft Sciences vs. ELT | | Hard Sciences vs. ELT | | Soft Sciences vs. Hard Sciences | |
|---|---|---|---|---|---|---|
| The beginning of the | 73 | 88 | 0 | | 0 | |
| It is possible to | 63 | 70 | 0 | | 0 | |
| The fact that the | 51 | 43 | 0 | | 0 | |
| In line with the | 47 | 121 | 0 | | 0 | |
| It is seen that | 45 | 55 | 0 | | 0 | |
| In accordance with the | 44 | 52 | 0 | | 0 | |
| In the form of | 43 | 44 | 0 | | 0 | |
| In addition to the | 42 | 65 | 0 | | 0 | |
| It can be said | 41 | 44 | 0 | | 0 | |
| When it comes to | 41 | 53 | 0 | | 0 | |
| As one of the | 0 | | 49 | 61 | 0 | |
| In the case of | 0 | | 0 | | 51 | 64 |

In addition, among the 4-word lexical bundles retrieved, there are also lexical bundles included in only one of these 3 sub-corpora. Table 5 shows that 56 out of these 94 lexical bundles are just peculiar to ELT, 6 to soft sciences and 9 to hard sciences.

**Table 5.** Frequency of Lexical Bundles peculiar to each sub-corpus

| Lexical Bundles | Hard Sciences | Soft Sciences | ELT |
|---|---|---|---|
| As shown in figure | 105 | - | - |
| The effect of the | 98 | - | - |
| With the help of | 74 | - | - |
| As shown in table | 55 | - | - |
| It can be seen | 51 | - | - |
| The performance of the | 46 | - | - |
| Be considered as a | 45 | - | - |
| Are shown in figure | 43 | - | - |
| The details of the | 42 | - | - |
| On the one hand | - | 60 | - |
| A part of the | - | 55 | - |
| In the context of | - | 51 | - |
| Within the scope of | - | 43 | - |
| As well as the | - | 42 | - |
| In the name of | - | 42 | - |
| In the present study | - | - | 157 |
| The findings of the | - | - | 156 |
| In the current study | - | - | 151 |
| Of the present study | - | - | 108 |
| It was found that | - | - | 103 |
| In terms of their | - | - | 100 |
| To be able to | - | - | 99 |
| It is observed that | - | - | 95 |
| Was found to be | - | - | 95 |
| Phase of the study | - | - | 89 |
| Were found to be | - | - | 86 |
| End of the study | - | - | 85 |
| It was observed that | - | - | 82 |
| Of the current study | - | - | 82 |
| At the beginning of | - | - | 80 |
| Of the study the | - | - | 79 |
| Most of the time | - | - | 76 |
| The participants of the | - | - | 75 |
| That there is a | - | - | 74 |
| In face to face | - | - | 73 |
| The analysis of the | - | - | 71 |
| The participants in the | - | - | 70 |
| As a foreign language | - | - | 66 |
| The results showed that | - | - | 64 |
| Findings of the study | - | - | 63 |
| English as a foreign | - | - | 60 |
| It was seen that | - | - | 59 |
| In a similar vein | - | - | 58 |
| All of the participants | - | - | 54 |
| In line with this | - | - | 53 |
| That there was a | - | - | 53 |
| In other words the | - | - | 50 |
| In the field of | - | - | 50 |
| As a part of | - | - | 49 |
| In a foreign language | - | - | 49 |
| Can be seen in | - | - | 48 |
| It is clear that | - | - | 47 |
| The second phase of | - | - | 47 |
| In relation to the | - | - | 46 |
| Participants of the study | - | - | 46 |

| | | | |
|---|---|---|---|
| Results of the study | - | - | 46 |
| As shown in the | - | - | 45 |
| Second phase of the | - | - | 45 |
| With the findings of | - | - | 45 |
| Of the participants in | - | - | 44 |
| The current study the | - | - | 44 |
| Can be said that | - | - | 43 |
| From time to time | - | - | 43 |
| That most of the | - | - | 43 |
| Thus it can be | - | - | 42 |
| In the target language | - | - | 41 |
| Most of the participants | - | - | 41 |
| In the process of | - | - | 40 |
| Of the students were | - | - | 40 |
| They were asked to | - | - | 40 |
| With respect to the | - | - | 40 |

The findings so far reveal that ELT dissertations contain significantly higher numbers of lexical bundles compared to dissertations in the fields of soft and hard sciences, with ELT dissertations having three to four times more lexical bundles. Interestingly, when it comes to the variety of lexical bundle types, the academic fields of soft and hard sciences exhibit similar patterns, implying that there is little variation in the usage of 4-word lexical bundles between these academic disciplines.

## Structural Analysis of Lexical Bundles

The findings towards the structural analysis of lexical bundles are shown in Figure 1. It seems that the expressions formed with noun and prepositional phrases are the most frequently used lexical bundles in each sub-corpus. Lexical bundles which appear within prepositional phrases outnumber even those within noun phrases in each sub-corpus.

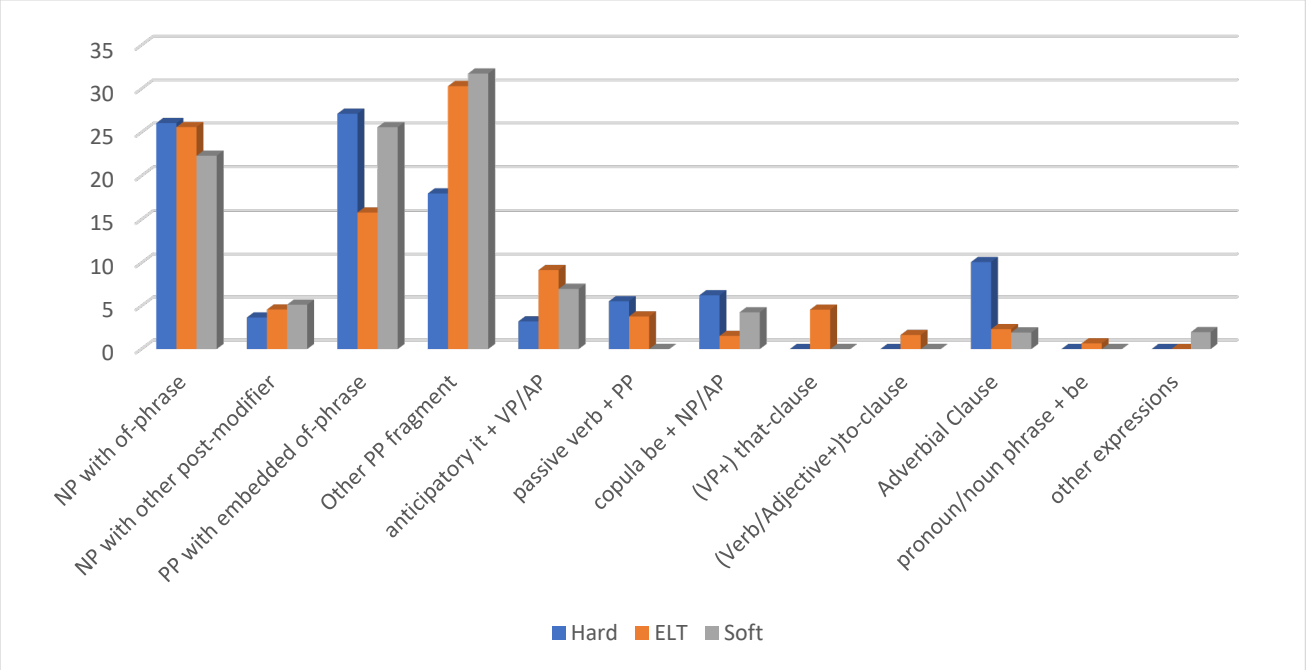


**Figure 1.** Structural Analysis of Lexical Bundles in Each Sub-corpus (%)

There are not any statistically significant differences among sub-corpora with regard to the use of lexical bundles within Noun Phrase with of-phrase fragment (7,06 LL value for ELT-Soft Science corpora; 0,10 LL value for ELT-Hard science corpora; 5,30 LL value for Soft-Hard Sciences p > ,0001 for each). Similarly, the sub-corpora

do not statistically differentiate in terms of the use of lexical bundles formed within Noun phrase with other post-modifier fragment (1,04 LL value for ELT-Soft Science; 2,60 LL value for ELT-Hard Science; 4,61 LL value for Soft-Hard Science p > ,0001). However, the sub-corpora differ in terms of the use of lexical bundles in the grammatical structure of prepositional phrases. For example, the authors of ELT dissertations significantly underuse lexical bundles (15,80%) within Prepositional phrases with embedded of-phrase fragment compared to the authors of Soft Science dissertations (25,61%) (76,93 LL value p < ,0001) and Hard Science dissertations (27,15%) (80,44 LL value p < ,0001). However, the underuse of lexical bundles within Prepositional phrases with embedded of-phrase fragment in Soft Science dissertations (25,61%) compared to Hard Science dissertations (27,15%) does not cause a statistically significant difference (0,81 LL value p > ,0001). On the other hand, although the sub-corpora ELT (30,31%) and Soft Science (31,75%) do not significantly differ (1,02 LL value p > ,0001) in terms of the use of lexical bundles within other prepositional phrase fragment, dramatic decrease of lexical bundles formed with other prepositional phrase fragment in Hard Science sub-corpus (17,99%) compared to the other two sub-corpora (ELT 30,31%; Soft Science 31,75%) causes a significant difference between Hard Science and ELT sub-corpora (76,06 LL value p > ,0001), and between Hard Science and Soft Science sub-corpora (69,16 LL value p > ,0001). Similarly, as in the use of lexical bundles within the structure of other prepositional phrase fragment, although ELT and Soft Science sub-corpora do not statistically differ (9,14 LL value p > ,0001), the dramatic underuse of lexical bundles within anticipatory it + VP/AP in Hard Science sub-corpus causes a significant difference between both ELT (9,17%) / Hard Science (3,22%) (68,51 LL value p < ,0001), and Soft Science (6,99%) / Hard Science (3,22%) sub-corpora (25,44 LL value p < ,0001).

## Functional Analysis of Lexical Bundles

Figure 2 indicates the findings of the functional analysis of lexical bundles in each sub-corpus. It is clearly shown that referential expressions appear as the most frequently attributed function to lexical bundles in each of the three sub-corpora. While discourse organizers are the second most attributed function to lexical bundles in Soft Science and Hard Science dissertations, referential expressions are followed by stance markers in ELT dissertations.

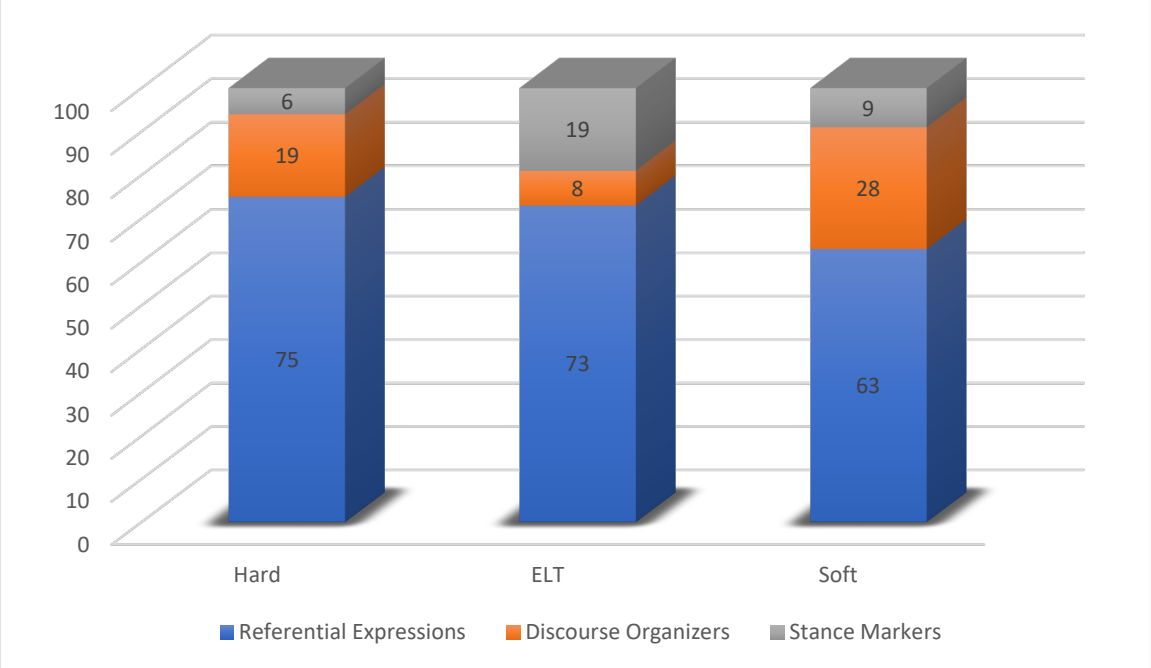


**Figure 2.** Functional analysis of lexical bundles in each sub-corpus (%)

A slight overuse of referential lexical bundles in Hard Science (75,44%) compared to ELT (73,31%) does not cause a significant difference between these two sub-corpora (0,77 LL value p > ,0001); however, the apparent underuse of referential lexical bundles in Soft Science dissertations (62,57%) significantly differentiates this sub-

corpus from both ELT (26,44 LL value p < ,0001) and Hard Science (21,95 LL value p < ,0001) sub-corpora. With regard to the use of discourse organizer lexical bundles, the overuse of lexical bundles as discourse organizers in Soft Science sub-corpus (28,05%) makes it significantly different from ELT (8,26%) (396,96 LL value p < , 0001) and Hard Science (18,5%) (35,52 LL value p < ,0001) sub-corpora. The underuse of lexical bundles as discourse organizers in ELT sub-corpus (8,26%) compared to Hard Science sub-corpus (18,5%) is also meaningful (109,47 LL value p < ,0001). In addition, although Soft (9,38%) and Hard Sciences (6,06%) do not differ in terms of the use of lexical bundles as stance markers (12,94 LL value p > ,0001), ELT dissertations (18,43%) differ from both fields of science (Soft Science 89,92 LL value; Hard Science 149,67 LL value p < ,0001).

## Discussion

The comparison of disciplines sorted out based on academic fields such as soft and hard sciences reveal that PhD dissertations of two different academic fields do not differ in terms of the token frequency of 4-word lexical bundles. When it comes to the comparison of major-English and non-English-major PhD dissertations, it is found that ELT dissertations significantly differentiate from the dissertations written in both soft and hard sciences. Dissertations written by major-English PhD candidates contain dramatically higher number of 4-word lexical bundles compared to those authored by non-English major PhD candidates. In the current study, although ELT dissertations contain more lexical bundles (0,68%) than the other two sub-corpora (0,23% Soft vs. 0,22% Hard), the use of 4-word lexical bundles is quite low considering their ratios in the analyzed data. This finding is contrary to previous studies which have suggested that non-native speakers of English use lexical bundles frequently even more than native speakers of English do (Bal-Gezegin, 2019; Güngör & Uysal, 2016; Öztürk & Durmuşoğlu-Köse, 2016). In the current study, while 78 different 4-word lexical bundles (0,68%) were found in ELT dissertations, this number decreased to 22 (0,22%) and 28 (0,23%) in hard and soft sciences, respectively. However, while there are 98 lexical bundles in the research articles of Turkish L2 speakers of English (Güngör & Uysal, 2016), 125 lexical bundles were found in PhD dissertations and master's theses of Turkish postgraduate students (Öztürk & Durmuşoğlu-Köse, 2016). Bal-Gezegin (2019) is another study in which the number of lexical bundles (99 4-word lexical bundles) in research articles from 6 disciplines authored by L1 Turkish scholars is higher than that of the current study. A possible explanation for this difference between the results might be that the cut-off point for the retrieval of the 4-word lexical bundles in each of these studies was determined in a different way. While the cut-off points 20 occurrences in Bal-Gezegin (2019) and Güngör and Uysal (2016), 25 occurrences in Öztürk and Durmuşoğlu-Köse (2016) may have resulted in a larger number of lexical bundles in these studies, setting a high cut-off point with a relatively small corpus, 40 occurrences of lexical bundles in dissertations across the sub-corpora, may have retrieved fewer lexical bundles in the present study. Another possible explanation for the relatively lower number of lexical bundles in the current study might be that even the distribution criteria to avoid the retrieval of individual idiosyncratic use of lexical bundles were different across these studies. Lexical bundles which appear at least 10% (Güngör & Uysal, 2016) and at least 5 of the total number of texts (Bal-Gezegin, 2019; Öztürk & Durmuşoğlu-Köse, 2016) were retrieved in these studies. However, in the current study, the lexical bundles which appear at least 20% of the texts were retrieved. It seems that the lower the cut-off point is set to avoid individual idiosyncratic use of lexical bundles, the more lexical bundles are obtained as target data. The findings of the current study related to the lower number of lexical bundles attained at the end of the frequency analysis are also contrary to previous studies (Bao & Liu, 2022) which have suggested, based on the token frequency of lexical bundles instead of the number of lexical bundles attained as in aforementioned studies, that non-native speakers of English use substantial amount of lexical bundles, even much more than native speakers of English while writing the abstracts of their dissertations. Bao and Liu (2022) find that dissertation abstracts of Chinese postgraduate students include 274 3-word lexical bundles with a total amount of 20,154 tokens, which equal to 3,28% of the whole corpus. However, total amount of 6,050 tokens of lexical bundles even in ELT dissertations in the current study equals to 0,68 normalized occurrences. The distribution criterion in Bao and Liu (2022) is also quite low, which equals to 2% of the texts, namely, 14 out of 700 dissertations. It can be inferred that although the number of lexical bundles attained increases as the cut-off point decreases, the risk of the retrieval of individual idiosyncratic use of lexical bundles also increases. The analysis of both 3-word lexical bundles and only the abstract sections of PhD dissertations in Bao and Liu's study (2022), opposed to analyzing the 4-word lexical

bundles in the whole PhD dissertations in the current study, may also be one of the potential reasons for the difference between the two studies.

One-by-one analysis of the lexical bundles in the present study indicates both similarities and differences with previous studies. For example, *on the other hand*, which is the most frequent type in each of the sub-corpora in the current study, is also among the most frequent 4-word lexical bundles in a number of studies (Bal-Gezegin, 2019; Biber et al., 1999; Hyland, 2008b; Perez-Llantada, 2014). However, it is somewhat surprising that any occurrences of *in the case of* is noted in the ELT sub-corpus although it is frequently available in Hard and Soft Science sub-corpora. This finding contrasts with a number of previous studies (Bal-Gezegin, 2019; Biber et al., 1999; Cortes, 2008) in which *in the case of* is really abundant. Although both the present study and the aforementioned studies examine lexical bundles in academic prose, the reason why *in the case of* is not found in the ELT dissertations may be discipline-specific. The corpora formed of with the inclusion of a number of disciplines except ELT field in Bal-Gezegin (2019), Biber et al. (1999), Cortes (2008), even Hard and Soft Science sub-corpora in the current study imply that some lexical bundles, by their nature, are not suitable for use in all disciplines.

Almost one-third of the lexical bundles in the soft science sub-corpus are the same as the lexical bundles in the ELT sub-corpus, while only one of them in the hard science sub-corpus is the same as the ELT sub-corpus. This rather interesting finding could be attributed to the fact that foreign language teaching, even if it is a scientific discipline among educational sciences, much resembles soft science dealing with human and animal behavior.

With regard to the structural analysis of lexical bundles in the current study, it seems that lexical bundles occurred within noun phrases are almost one-third of the total number of lexical bundles in each of the sub-corpora. On the other hand, lexical bundles within the grammatical structure of prepositional phrases are almost half of the total number of lexical bundles in ELT and Hard science sub-corpora, even more than half in soft science sub-corpus. These results corroborate the findings of a great deal of the previous work. For example, Cortes (2008), who investigates the use of lexical bundles in academic history writing in both English and Spanish, finds that lexical bundles are mostly included in prepositional phrases and noun phrases. Allen (2010), comparing learners' lexical bundles use to scholars with published research articles and to scholars who are native speakers of English, finds that 'noun phrases with of-phrase fragment' is the most frequently encountered linguistic structure in which lexical bundles appear. In a similar vein, Hyland (2008), analyzing PhD dissertations, master's theses and research articles from 4 different disciplines, finds that 'noun phrase with of-phrase fragment' is the most typical grammatical structure which contains 4-word lexical bundles in it.

As for the functional analysis of lexical bundles, the overwhelming majority of lexical bundles are the ones used to be referential expressions in the current study. The ELT sub-corpus significantly differs from other corpora with relatively less use of stance markers and relatively higher use of discourse organizers. Similar findings were also reported by Bal-Gezegin (2019) and Cortes (2008). Bal-Gezegin (2019), analyzing research articles published by Turkish authors, finds that while 75% of the lexical bundles function to be referential expressions, 8% of them function to be stance markers to denote authorial stance. Similarly, Cortes (2008) who compares history research articles written in English and Spanish finds that the lexical bundles available in research articles in both languages function to be referential expressions. It can be inferred that the authors in academic prose most frequently use lexical bundles to refer to either all, some, or one of the concrete or abstract elements in the text.

## Conclusion

The present study reveals that PhD dissertations sorted out based on their academic fields such as soft and hard sciences do not differentiate from each other with regard to the use of lexical bundles. However, ELT dissertations, the representative of English-major disciplines, significantly differ from hard and soft sciences in terms of the use of lexical bundles. On the other hand, the structural and functional analyses of the lexical bundles in these three sub-corpora reveal that noun phrases and verb phrases are the syntactic structures in which the lexical bundles most frequently appear in each of the sub-corpora. Furthermore, the lexical bundles in these sub-corpora mostly function to be referential expressions. The evidence from the present study implies

that although the lexical bundles in soft and hard science dissertations are structurally and functionally similar to some extent, they have lagged far behind the lexical bundles in the ELT dissertations in terms of type/token frequency.

The main reason for using dramatically more lexical bundles in the English-major discipline compared to non-English major disciplines might be the level of foreign language proficiency. It might be assumed that postgraduate student authors of ELT dissertations are highly proficient in English. Byrd and Coxhead (2010) emphasize the importance of lexical bundles "to develop both fluency and appropriate usage for particular settings" (p.32). In academic prose as a particular setting, increasing language proficiency may have allowed both the development of academic writing skills and the use of more appropriate expressions for academic writing. In other words, it can be inferred that increasing language proficiency has a positive effect on the adoption of academic writing conventions, among which precise word and word combination choice is indispensable. Raising awareness on lexical bundles and teaching them explicitly might come to the forefront as a pedagogical implication which will undoubtedly contribute to the academic writing skills of the authors. Concentrating on the explicit teaching of lexical bundles will both pave the way for processing words as a group instead of individually in the mind and allow students to speak more fluently.

The present study reveals that ELT sub-corpus as the representative of English-major disciplines obviously includes much more lexical bundles than non-English major disciplines. Future studies might explore disciplines in different academic fields more specifically, as in Cortes (2004, 2008) examining academic studies in history or Hyland (2008b) analyzing lexical bundles in academic studies in electrical engineering, business studies, applied linguistics, and microbiology. In the current study, lexical bundles in the PhD dissertations of postgraduate students, who were about to be an independent researcher and expert writer, were analyzed. Further research may also focus on examining master's and doctoral theses by the same authors and determining whether gaining experience in a particular field of science leads to a change in the use of lexical bundles.

On the other hand, in the current study, it can be considered as a limitation that while including the dissertations as data in the research, no attention was paid to whether their methods were theoretical or empirical. Theoretical or empirical studies, or even more specifically quantitative and qualitative research, may necessarily contain different formulaic expressions. It might be more plausible to consider the division of dissertations, organized according to their academic disciplines, into subgroups based on the specific research methodology utilized to potentially enhance the findings of future studies in a more detailed way.

## References


Allen, D. (2010) Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. *Komaba Journal of English Education,* 1, 105-127.

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

Anthony, L. (2022). AntConc (Version 4.2.0) [Computer Software]. Tokyo, Japan: Waseda University. Available from 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.

Bao, K., & Liu, M. (2022). A corpus study of lexical bundles used differently in dissertations abstracts produced by Chinese and American PhD students of Linguistics. *Frontiers in Psychology*, 13, 1-13. https://doi.org/10.3389/fpsyg.2022.893773

Biber, D., & Barbieri, F. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes*, 26(*3*), 263–286.

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at . . .: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, 25(*3*), 371-405.

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

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. (2008). A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora*, 3(*1*), 43-57. https://doi.org/10.3366/E1749503208000063

Dahunsi, T. N., & Ewata, T. O. (2022). An exploration of the structural and colligational characteristics of lexical bundles in L1–L2 corpora for English language teaching. *Language Teaching Research*, *0*(0). https://doi.org/10.1177/13621688211066572 (Epub ahead of print)

Erman, B., & Warren, B. (2000). The idiom principle and the open-choice principle. *Text*, 20, 29-62. https://doi.org/10.1515/text.1.2000.20.1.29

Granger, S. (1998). Prefabricated patterns in advanced EFL writing. Collocations and formulae. In A. P. 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.

Hsu, W. (2014). The most frequent opaque formulaic sequences in English-medium college textbooks. *System*, 47, 146-161. http://dx.doi.org/10.1016/j.system.2014.10.001

Hyland, K. (2008). 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(*1*), 41-62.

Hyland, K. (2012). Bundles in academic discourse. *Annual Review of Applied Linguistics*, 32, 150-169.

Kecskes, I. (2007). Formulaic language in English lingua franca. In I. Kecskes, & L. Horn (Eds.), *Exploration in pragmatics: Linguistic, cognitive and intercultural aspects* (pp.191-218). Berlin/New York: De Gruyter Mouton. https://doi.org/10.1515/9783110198843

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. NewYork: Newbury House.

Ortaçtepe, D. (2013). Formulaic language and conceptual socialization: The Route to becoming nativelike in L2. *System*, 14, 852-865. http://dx.doi.org/10.1016/j.system.2013.08.006

Öztürk, Y., & Köse, G. D. (2016). Turkish and native english academic writers' use of lexical bundles. *Journal of Language and Linguistic Studies*, 12(*1*), 149-165.

Pérez-Llantada, C. (2014). Formulaic language in L1 and L2 expert academic writing: Convergent and divergent usage. *Journal of English for Academic Purposes*, 14, 84-94. http://dx.doi.org/10.1016/j.jeap.2014.01.002

Schmitt, N. (2000). *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.

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

Rayson, P., Berridge, D., & Francis, B. (2004). Extending the Cochran rule for the comparison of word frequencies between corpora. In G. Purnelle, C. Fairon and A. Dister (Eds.). *Le Poids des Mots. Proceedings*

*of the 7th International Conference on Statistical Analysis of Textual Data (JADT 2004)* (pp. 926-936). Louvain: Presses Universitaires de Louvain.

Shin, D., & Nation, P. (2007). Beyond single words: the most frequent collocations in spoken English. *ELT Journal,* 62(*4*), 339–348. https://doi.org/10.1093/elt/ccm091

Thornbury, S. (2002). *How to teach vocabulary*. Essex: Pearson Education Limited.

Üstünbaş, Ü. & Ortaçtepe, D. (2016). EFL learners' use of formulaic language in oral assessments: A study on fluency and proficiency. *Hacettepe University Journal of Education*. 31(*3*), 578-592.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

## GENİŞLETİLMİŞ ÖZET

Kelime bilgisi, Nation'ın (1990) da ele aldığı gibi, kelimelerin eşdizimliliklerinin ve çağrışımlarının kelime dağarcığının tek kelimeden daha fazlası olduğu gerçeğine işaret ettiği, sözcüksel bilginin birçok farklı yönü üzerinde ustalık kazanmayı gerektiren karmaşık bir konudur. Schmitt (2000) ve Schmitt ve Carter (2004), kelimelerin insan zihninde tek tek değil gruplar halinde organize edildiğinden bahsetmektedir. Byrd ve Coxhead (2010) ise derlem dilbilimin dilde sık sık tekrarlayan ve konuşmanın akıcı olması ve kulağa normal gelmesi açısından vazgeçilmez olan hem kalıplaşmış hem de yarı kalıplaşmış ifadelerin ortaya çıkarılmasına yardımcı olduğunu belirtmektedir. Söz öbekleri uzun zamandır geniş bir araştırma yelpazesinde büyük ilgi konusu olmuştur. Söz öbekleri ile ilgili önceki çalışmalar, bu kalıplaşmış ifadeleri birçok farklı yönden ele almıştır. Örneğin, söz öbeklerinin sıklıklarını belirlemek için hem mega derlemler hem de belli bir amaç için oluşturulan derlemler incelenmiştir. Araştırma makaleleri, doktora tezleri, yüksek lisans tezleri, sınıf öğretimi söylemi, ders kitapları ve sözlü yeterlilik sınavları söz öbekleri çalışmalarında veri olarak karşımıza çıkmaktadır. Öte yandan söz öbeklerinin sıklığı tek başına çok şey ifade etse de pedagojik çıkarımlar içermemekte ve sınıf ortamında gerçekleştirilen dil öğretimi çalışmalarına fazla bir katkı sağlamamaktadır. Bu nedenle sözcük yığınlarının yapısal ve işlevsel özelliklerini de inceleyen çok sayıda çalışma mevcuttur. İlgili alan yazın aynı zamanda anadili İngilizce olmayan kişiler ile anadili İngilizce olan kişiler arasındaki söz öbeklerinin kullanımını karşılaştıran çalışmalar da içermektedir. Mevcut çalışma, anadili İngilizce olmayan lisansüstü öğrenciler tarafından yazılan doktora tezlerini bilim alanlarına göre karşılaştırarak yukarıda bahsedilen çalışmaların bulgularını genişletmeye çalışmaktadır. Doktora tezlerinde tekrar eden söz öbeklerini tespit etmek ve yapısal ve işlevsel analizlerini yapabilmek için çeşitli disiplinler, fen bilimleri ve sosyal bilimler gibi bilimsel alanlara dayalı olarak iki ayrı gruba ayrılmıştır. Ayrıca, uzmanlık alanı İngilizce olan ve uzmanlık alanı İngilizce olmayan disiplinler birbiriyle karşılaştırılmıştır.

Mevcut çalışma, küçük ölçekli üç alt derlemden oluşan bir derlemi incelemektedir. Anadili İngilizce olmayan yazarlar tarafından İngilizce olarak yazılmış doktora tezlerinde tekrar eden söz yığınlarını belirlemek için üç farklı alt derlem oluşturulmuştur. Sosyal bilimler alt derlemi 16 adet tezin dahil olduğu toplam 922,005 kelimeden oluşurken, fen bilimleri alt derlemi 723,082 kelimelik toplam 16 tezden oluşmaktadır. İngiliz Dili Eğitimi bilim dalında yazılmış 887,920 kelimeden oluşan 16 tez üçüncü alt derlemi oluşturmaktadır. Bahsedilen 3 alt derlem, söz öbeklerinin sayıları, ve mevcut söz öbeklerinin yapısal ve işlevsel özellikleri açısından birbirleriyle karşılaştırılmıştır. Bulgular, uzmanlığı İngilizce olan disiplinleri temsil eden İngiliz Dili Eğitimi doktora tezlerinin, sosyal ve fen bilimleri alanlarında yazılan tezlerden sırasıyla üç ve dört kat daha fazla söz öbeğine sahip olduğunu ortaya koymaktadır. Bununla birlikte, sosyal ve fen bilimleri alanlarında yazılan doktora tezleri hemen hemen aynı miktarda söz öbeği içermektedir, bu da sosyal ve fen bilimlerinin 4 kelimelik söz öbeklerinin kullanımında değişiklik göstermediğine işaret etmektedir. Söz öbeklerinin yapısal analizi, söz öbeklerinin en sık isim tümceleri ve edat tümceleri içerisinde yer aldıkları sonucunu ortaya çıkarmıştır. Söz öbeklerinin işlevsel dağılımıyla ilgili ise bulgular, her bir alt derlemdeki çok sayıda söz öbeğinin gönderge ifadeleri olduğunu göstermektedir.