REPUBLIC OF TURKEY
ÇUKUROVA UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING

MULTIWORD UNITS IN ACADEMIC WRITINGS BY NATIVE SPEAKERS OF ENGLISH AND TURKISH EFL LEARNERS: A LEXICAL BUNDLES APPROACH

Sibel AYBEK

DOCTOR OF PHILOSOPHY

ADANA / 2023

REPUBLIC OF TURKEY
ÇUKUROVA UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING

MULTIWORD UNITS IN ACADEMIC WRITINGS BY NATIVE SPEAKERS OF ENGLISH AND TURKISH EFL LEARNERS: A LEXICAL BUNDLES APPROACH

Sibel AYBEK

Supervisor : Prof. Dr. Cem CAN
Member of Examining Committee: Prof. Dr. Zuhal OKAN
Member of Examining Committee: Doç. Dr. Mustafa MAVAŞOĞLU
Member of Examining Committee: Prof. Dr. Türkay BULUT
Member of Examining Committee: Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

DOCTOR OF PHILOSOPHY

ADANA / 2023

To Çukurova University Institute of Social Sciences;

We certify that this thesis is satisfactory for the award of the degree of Doctor of Philosophy in the Department of English Language Teaching.

Supervisor: Prof. Dr. Cem CAN

Member of Examining Committee: Prof. Dr. Zuhal OKAN

Member of Examining Committee: Doç. Dr. Mustafa MAVAŞOĞLU

Member of Examining Committee: Prof. Dr. Türkay BULUT

Member of Examining Committee: Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

ONAY

I certify that this dissertation confirms to the formal standards of the Institute of Social Sciences. …/…/2023

Prof. Dr. Serap ÇABUK
Director of the Institute

ETİK BEYANI

Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmasında;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmasında yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde ve ortaya çıkan sonuçlarda herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim. …. / …. / 2023

Sibel AYBEK

ÖZET

İNGİLİZCE'Yİ ANA DİLİ OLARAK KONUŞANLAR VE YABANCI DİL OLARAK İNGİLİZCE ÖĞRENEN TÜRK ÖĞRENCİLERİN AKADEMİK YAZI DİLLERİNDE ÇOK SÖZCÜKLÜ BİRİMLER: SÖZCÜK ÖBEĞİ TEMELLİ BİR ÇALIŞMA

Sibel AYBEK

Doktora Tezi, İngiliz Dili Eğitimi Ana Bilim Dalı
Danışman: Prof. Dr. Cem CAN
Haziran 2023, 245 sayfa

Derlem dilbilimi kapsamında yapılan çalışmalar, özgün dil kullanımının, kalıp sözler olarak da bilinen tekrarlanan ifadeleri içerdiğini açıkça göstermiştir (Byrd & Coxhead, 2010). Bu ifadeler "hem sözlü hem de yazılı anlatımların temel yapı taşları" olarak tanımlanabilecek çok sözcüklü birimler olarak tanımlanabilir (Biber & Barbieri, 2007, s. 263). Kalıplaşmış bu diziler, sözlü ve yazılı dilin büyük bir kısmını oluşturur (Schmitt & Carter, 2004) ve "dil gelişiminde bir basamak taşı" işlevini görür (Chenu & Jisa, 2009, s. 27). Kalıp dizilerin önemli bir türü olan sözcük öbekleri, "deyimleşme ve yapısal durumlarına bakılmaksızın yinelenen ifadelerdir (Biber ve diğerleri, 1999, s. 990)." Bu çalışma, Uluslararası İngilizce Öğrenen Derlemi'nin (ICLE) bir alt derlemi olan Türk Uluslararası İngilizce Öğrenen Derleminde (TICLE) ve Louvain Anadil Olarak İngilizce Deneme Derleminde (LOCNESS) en sık kullanılan 3 ve 4 sözcüklü sözcük öbeklerinin kullanımını incelemektedir. TICLE derlemi, İngilizceyi Yabancı Dil olarak Öğrenen Türk öğrencilerin yazılı denemelerini içerirken, LOCNESS derlemi Amerikan öğrencilerin yazılı denemelerini içermektedir (Granger ve diğerleri, 2009). Çalışma, derlem tabanlı yaklaşımı (Sinclair, 1987) kullanmaktadır, aynı zamanda sıklık tabanlı yaklaşım (Nesselhauf, 2004) olarak da adlandırılır. En sık kullanılan diziler, anlatımdaki yapıları ve işlevleri çerçevesinde Biber ve diğerlerinin (2003; 2004) sözcük öbeklerinin yapısal ve işlevsel sınıflandırmasına göre incelenmiştir. Bu yapılar arasında Türk öğrenciler ve ana dil olarak İngilizce konuşanlar arasında betimsel karşılaştırmalı bir çözümleme yapılmıştır. Son olarak, öğrencilerin çok sözcüklü öbek seçimlerinin en belirgin özelliklerini ortaya çıkarmak amacıyla her iki derlemdeki çok sözcüklü ifadelerin

anlambilimsel alanları, Wmatrix (Versiyon 4) anlamsal çözümleme arayüzünü kullanarak UCREL Anlamsal Çözümleme Sistemi (USAS) (Rayson, 2008) anlamsal etiket setleri kullanılarak ortaya çıkarılmıştır. Çalışma, Türk öğrencilerin eylem öbeği bölümlemelerini aşırı kullanırken, ad ve ilgeç öbeklerinin ana dil olarak İngilizce konuşanlarla karşılaştırıldığında yeterince kullanmadıkları görülmüştür. İşlevsel sınıflandırma açısından, Türk öğrenciler daha az işlevsel öbek üretse de tekrar eden öbeklerin sınırlı bir kümesini fazla kullanmaktadırlar. Log-olabilirlik sonuçları da TICLE derleminde LOCNESS derlemine oranla sözcük çeşitliliği ve zenginliğinin daha az olduğunu göstermektedir. Ayrıca, bu anlatımların anlambilimsel çözümlemesi, her iki derlemde de benzersiz şekilde kullanılan hem ortak hem de özgül alanları ortaya çıkarmaktadır. Çalışma, Türk öğrencilerin yabancı dil olarak İngilizce öğrenmelerinde sözcük demeti kullanımına ilişkin bilgi birikimine katkıda bulunmakta ve İkinci Dil Edinimi (SLA) araştırmacıları, İngilizce öğretmenleri, eğitmenler ve materyal geliştiriciler ve yayıncılar için yararlı içgöçüler sunmaktadır.

Anahtar kelimeler: Derlem dilbilimi, öğrenen derlemi, çok sözcüklü birimler, sözcük öbekleri, yabancı dil olarak İngilizce öğrenen Türk öğrenciler

ABSTRACT

MULTIWORD UNITS IN ACADEMIC WRITINGS BY NATIVE SPEAKERS OF ENGLISH AND TURKISH EFL LEARNERS: A LEXICAL BUNDLES APPROACH

Sibel AYBEK

Ph.D. Thesis, Department of English Language Teaching
Supervisor: Prof. Dr. Cem CAN
June 2023, 245 pages

Corpus linguistics has clearly demonstrated that authentic language use consists of repeated expressions, namely formulaic expressions (Byrd & Coxhead, 2010). These expressions, which can also be described as multiword units, serve as essential "building blocks of discourse in both spoken and written registers" (Biber & Barbieri, 2007, p. 263). Formulaic sequences constitute a large portion of oral and written language production (Schmitt & Carter, 2004) and "provide a steppingstone into language development" (Chenu & Jisa, 2009, p. 27). Being an important type of formulaic sequences, lexical bundles are "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber et al., 1999, p. 990). The current study investigates the use of the most frequent 3- and 4-word lexical bundles in the Turkish International Corpus of Learner English (TICLE) which is sub-corpus of the International Corpus of Learner English (ICLE) and Louvain Corpus of Native English Essays (LOCNESS). The TICLE corpus contains written argumentative essays by Turkish EFL learners, while the LOCNESS corpus comprises American students' written essays (Granger et al., 2009). The study employs the corpus-driven approach (Sinclair, 1987), also referred to as frequency-based approach (Nesselhauf, 2004). The most frequent expressions have been classified according to their structures and functions in the discourse using Biber et al.'s (2003; 2004) structural and functional taxonomy of lexical bundles framework. An interpretative contrastive analysis of these expressions has been conducted between Turkish EFL learners and native English speakers. And lastly, semantic domains of the multiword expressions in both corpora have been revealed using UCREL Semantic Analysis System (USAS) semantic tag sets utilizing Wmatrix (Version 4) semantic

analysis platform (Rayson, 2008) to identify the most salient features of the multiword choices of learners. The study concludes that Turkish EFL learners overuse verb phrase fragments while they underuse the noun phrase and prepositional phrase fragments in relation to the native speakers of English. In terms of the functional classification, even though Turkish EFL learners produce less functional bundles, they overuse a limited set of repetitive bundles. Log-likelihood results also indicate a lack of lexical variety and richness in TICLE corpus in relative to the LOCNESS corpus. Additionally, semantic analysis of these phrases uncovers both common and specific domains uniquely used in both corpora. Current study adds our understanding of the use of English lexical bundles by Turkish EFL learners and offers valuable insights for SLA researchers, English language teachers, instructors and material developers and publishers.

Keywords: Corpus Linguistics, Learner Corpora, Multiword Units, Lexical Bundles, Turkish EFL Learners

ACKNOWLEDGEMENTS

My PhD journey has been a challenging path, characterized by numerous ups and downs, drawbacks, doubts and difficulty. At times, it seemed almost impossible to stay on course. As I reflect upon this dissertation, it has become an inseparable part of my identity, akin to a cherished child, embodying a significant facet of my intellectual growth and development. This work stands as a testament to the unwavering virtues of patience, persistence, laborious dedication, and profound sacrifices willingly made along the way.

I would like to express my heartfelt gratitude to all those who have supported me throughout this PhD journey. Without their unwavering support and guidance, this achievement would not have been possible.

First and foremost, I extend my deepest appreciation to my esteemed supervisor, Prof. Cem Can whose expertise, guidance, and invaluable feedback have shaped this dissertation and challenged me to push the boundaries of my research. His unwavering dedication and intellectual rigor to my academic and personal growth has been an inspiration, and I am truly grateful for the opportunity to work under his mentorship.

I am very grateful to dear Prof. Zuhal Okan for her dedication and support throughout my dissertation. Her commitment to academic excellence has been a constant source of motivation and inspiration and her valuable insights always helped me refine my ideas and contributed to shaping of this dissertation.

I would also like to express my profound appreciation to my examining committee, Prof. Türkay Bulut, Prof. Şehnaz Şahinkarakaş, and Assoc. Prof. Mustafa Mavaşoğlu, who has recently received the Order of Academic Merit (Ordre des Palmes Académiques) by the French Ministry of National Education and Youth upon his contributions to the French language and culture. Their valuable feedback, insightful comments, and unwavering commitments have been instrumental in shaping my research. Their constructive criticism, thoughtful suggestions, and insightful discussions have significantly enriched the depth and breadth of my work.

I would like to acknowledge Sylviane Granger and Susan Hunston for their generous contributions, stimulating discussions, insightful feedback, and patience for my never-ending questions. I also extend my gratitude to the corpus linguistics community, both online and offline, for their contributions. Engaging with fellow researchers, scholars and their publications have broadened my horizons and enriched my understanding of this dynamic field.

I owe special thanks to Res. Asst. Eylem Perihan Kibar for her support in any way and being a true friend every now and then. Each of our conversations, lunchbreaks, walking and hanging around the campus have been another source of inspiration for me. Sharing the office with you is also a privilege and I cannot thank enough for putting up with me, being with me at my worst and being an amazing company.

Also, my sincere appreciation and heartfelt gratitude extends to my dearest sister, Kadriye Aybek, who surprised me at the last minute by coming to my defense from Germany, provided immeasurable support and companionship. I thank her a million, her gesture has instilled in me an immense sense of happiness and confidence that words cannot adequately express. She has recently completed her bachelor's degree despite everything. Her perseverance and determination are a source of constant inspiration. I am immensely proud of her every now and then.

Last but not least, I also extend my gratitude to my esteemed honorary family, the Karaömerlioğlu (Selim, Nuray, Saim, and Cemal), and my dear friend Ramazan Özceviz, whose constant support, love, and encouragement have sustained me throughout this challenging journey. Their belief in my abilities and constant encouragement have given me the strength to persevere during moments of doubt and difficulty. Moreover, I would also like to thank to my beloved cat son, Çiko, who may be watching over me from wherever he is. I carry fond memories of him, and he is always in my thoughts.

To all those who have played a role, however big or small, in shaping this dissertation and supporting me along the way, I express my deepest gratitude. Collective contributions have transformed this research endeavor into a reality.

CONTENTS

Page

CHAPTER III

METHODOLOGY

CHAPTER IV

FINDINGS

CHAPTER V
DISCUSSION

CHAPTER VI
CONCLUSION AND SUGGESTIONS

LIST OF ABBREVIATIONS

AWL: Academic Word List

DC: Dependent Clause

DDL: Data-Driven Learning

EAP: English for Academic Purposes

EFL: English as Foreign Language

ELT: English Language Teaching

ESL: English as Second Language

F: Frequency

FLT: Foreign Language Teaching

FREQ: Frequency

ICLE: International Corpus of Learner English

KWIC: Key Word in Context

L1: First Language

L2: Second Language

LB: Lexical Bundle

LL: Log-Likelihood

LOCNESS: Louvain Corpus of Native English Essays

MWI: Multiword Items

N: Number

NP: Noun Phrase

NS: Native Speaker

NSS: Non-native Speaker

POS: Part-of-Speech

PP: Prepositional Phrase

T/t: Type/Token ratio

TICLE: Turkish International Corpus of Learner English

UCREL: University Centre for Computer Corpus Research on Language

USAS: UCREL Semantic Analysis System

VP: Verb Phrase

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF APPENDICES

Page

CHAPTER I

INTRODUCTION

1.1. Background to the Study

Scholars have long been interested in the use of recurrent patterns. Firth (1957) stated that "you shall know a word by the company it keeps" (p. 195) highlighting the significance of the patterns in the surrounding context to understand a word. Similarly, Hymes (1968) claimed that "a vast portion of verbal behavior … consists of recurrent patterns, of linguistic routines" (p. 126) by looking at the social functions of language. The status of multi-word expressions has also been examined by branches of lexicography (Moon, 1997). Since studies on multi-word sequences increased, their significance has also been revealed to a great extent. It has also been revealed that large part of the language use consists of these repeated expressions. We have only realized this particularly when corpus-based research has taken part in lexical studies. One of the earliest studies to mention this belongs to Altenberg (1998) who underlined that 80% of the words in the corpus are recurrent sequences. She confirmed Sinclair's study (1991) propounding the "idiom principle" that is though the phrases might appear to be analyzable into smaller segments, "a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices" (Sinclair, 1991, p. 110).

"Phraseology" (Granger & Meunier, 2008; Meunier & Granger, 2008) and "formulaic sequences/language" (Schmitt, 2004; Wray, 2002) are two umbrella terms that refer to various types of multi-word units. Corpus linguistics has clearly demonstrated that the authentic language use comprises repeated expressions (Byrd & Coxhead, 2010), namely formulaic expressions. These expressions occur together and could be referred to as multiword units that "are important building blocks of discourse in spoken and written registers" (Biber & Barbieri, 2007, p. 263). Various terms have been used to refer these multiword units such as 'clusters' (Hyland, 2008a; Schmitt, Grandage & Adolphs, 2004; also used in the corpus tool WordSmith (Scott, 2020)), 'recurrent word combinations' (Altenberg, 1998; De Cock, 2004), 'phrasicon' (De Cock, Granger, Leech, & McEnery, 1998), 'n-grams' (Stubbs, 2007a, 2007b) and 'lexical bundles' (Biber & Barbieri, 2007; Cortes, 2002). These terms—clusters, phrasicon, n-grams, recurrent word combinations, lexical bundles— refer to continuous word sequences generated by conducting a corpus-

driven process with specified frequency and distribution criteria. These recurrent sequences are fixed multi-word units which have pragmatic and/or discourse functions. This methodology is considered to be a frequency-based approach for determining phraseology as Granger & Paquot (2008) put forward.

Corpus linguistics provides ways to explore and analyze the language uses for learners by authentic language data for language learning and teaching through keyword lists, concordances, and word frequencies (Granger, 2002). Corpus linguistics is not only used in applied linguistics, speech research, grammar, sociolinguistics, and culture, but it is also used in areas such as semantics and lexical studies. Thanks to the lexical studies, corpus-informed dictionaries have been compiled and written with up-to-date information (McEnery& Wilson, 2001). The use of corpus for pedagogical purpose has enabled both teachers and learners to use corpus-based educational products such as dictionaries, textbooks, and software now more than ever (Gabrielatos, 2005).

1.2. Statement of the Problem

Several studies have been carried out to highlight the importance of the multiword units, one of which is by Altenberg (1998), who asserted that 80% of the words in the corpus formed the part of recurrent word combinations in London-Lund Corpus. Biber et al. (1999) also indicated that identifying and describing patterns of use through large-scale corpus studies can enlighten an aspect of language that is often ignored by noting as below:

> Grammar is not just a study of abstract classes and structures, but of particular words and their particular functions within those classes and functions. Such information is also important for the learner of English as a foreign language: producing natural, idiomatic English is not just a matter of constructing well-formed sentences, but of using well-tried lexical expressions in appropriate places (p. 989).

Similarly, Biber, Conrad and Cortes (2004) argue that everyday language use comprises multi-word prefabricated expressions (e.g., *if you see, in a nutshell, what I mean*) and language is not "strictly compositional" (p. 372). A multi-million-word corpus involving conversation and academic prose has been analyzed by Biber et al. (1999) in terms of lexical sequences. In their study, they investigated the uses of lexical bundles in different contexts and created a framework for the classification of lexical bundles (LBs

henceforth) (Biber, Conrad, & Cortez, 2004). Even though many studies have been carried out on formulaic language in the last decade, as Nekrasova (2009) claims, more study is needed to have a closer look at the structural and functional characteristics of lexical phrases.

Martinez & Schmitt (2012) argue that language production and processing mostly base upon a holistic retrieval of formulaic sequences as a single unit instead of a creative structure of utterances considering grammar rules (Wray, 2002). Therefore, these ready-made lexico-grammatical combinations in different language domains assist learners achieve "naturalness in language use" (Allen, 2009, p. 106). Likewise, producing a great range of formulaic sequences reveals the foreign language proficiency level of learners. The pivotal place that formulaic sequences take in second language acquisition (SLA) field enabled researchers to investigate their use, development, and learning in different fields like applied linguistics, corpus linguistics and psycholinguistics (Li & Schmitt, 2009; Reppen, 2009; Schmitt et al., 2004; Biber & Conrad, 2009; Crossley & Salsbury, 2011). This study aims to contribute the related studies on the formulaic sequences that are referred to as "lexical bundles" being a vital component of the fluent linguistic production (Hyland, 2008a) by adding frequency-driven approach methodology (Nesselhauf, 2004). And in this dissertation, multiword units, recurrent word combinations, multiword expressions and lexical phrases have been used interchangeably to refer to the lexical bundles.

While most of the studies focus on the academic written and spoken corpora of the English language (Cortes, 2002; Hyland, 2008a; Hyland, 2008b; Neely and Cortes, 2009; Bal, 2010), some of them center on teaching and textbook language (Biber, Conrad & Cortes, 2004), and some others deal with comparing the use of lexical bundles produced by both native and non-native speakers (NNS henceforth) of English (Cortes, 2004; Juknevičienė, 2011). In recent years, there have been some studies comparing the use of LBs of learners from different native language backgrounds such as Wei & Lei (2011), Adel & Erman (2012), Muşlu (2015), Chen and Baker (2016) and Shin (2018). More studies with learners from different backgrounds are needed to be able to generalize the production of the lexical bundles and have more insights about whether the usages are specific to different L1s of the learners or not. Hyland (2008b) and Granger (2002) assert that as more and detailed understanding of learner output are revealed, all aspects of pedagogy including curriculum designs, materials and teachers may be informed.

A massive amount of evidence has been provided for the inseparability of lexis

and grammar by recent research in corpus linguistics (Meunier & Granger, 2008; Biber, Conrad, & Cortes, 2004; Hunston & Francis, 2000; Hoey, 2005). Hence, this necessitates a more integrated perspective on the language by highlighting frequently used phrases that serve as significant meaning-carrying elements in the discourse according to Römer (2011). Moreover, studies reveal that productions from learners deviate from native speakers (NS henceforth), that's why, learners should be exposed multi-word items with Data-Driven Learning (DDL henceforth) exercises and corpus-derived teaching materials developed by teachers and corpus researchers.

Therefore, this study aims to contribute to the related studies by investigating structures and functions of the formulaic sequences in the written productions of Turkish EFL learners adopting a frequency perspective. Offering interpretations on how learners produce formulaic language in their writing through examining their written production provides insights to the body of research on the development of formulaic sequences by helping English language teachers, instructors and material developers understand the nature of these multi-word combinations.

1.3. Purpose of the Study and The Research Questions

The current study aims to investigate the use of 3- and 4-word lexical bundles in the Turkish International Corpus of Learner English (TICLE) learner corpus containing the written argumentative essays by Turkish EFL learners. The study particularly examined the most frequently used expressions, their structures, and functions in the written performances of Turkish university students. The purpose has been to provide some understanding of the use and differences of these units.

After revealing the structural and functional features of the lexical bundles of Turkish EFL learners, similarities and differences have been analyzed comparing the use of those sequences by American university students in argumentative essays. The native speaker corpus used in the present study is Louvain Corpus of Native English Essays (LOCNESS) which is a reference corpus containing American university students' written essays (Granger et al., 2009). The structure and function of the most frequent lexical bundles have been revealed according to Biber et al.'s (1999; 2003; 2004) structural and functional taxonomy of lexical bundles framework. Additionally, similarities and differences of the frequent 3- to 4-word bundles between Turkish EFL learners and native English speakers have been examined. A Log-likelihood statistical

analysis have been carried out in order to see the normalized frequencies between two corpora. And lastly, semantic domains of these MWUs have been studied utilizing Wmatrix (Version 4) semantic analysis platform (Rayson, 2008) to identify the most salient features of the multiword choices of learners. Even though previous studies examined bundles' functions in discourse, only few studies investigated the semantic domains of LBs as Cortes & Hardy (2013) discuss.

In this regard, research questions of the study include:

1. What are the most frequent 3- and 4-word lexical bundles found in argumentative essays of Turkish EFL learners in the Turkish International Corpus of Learner English (TICLE) and native speakers in the Louvain Corpus of Native English Essays (LOCNESS) corpora?
2. How are these lexical sequences found in TICLE and LOCNESS corpora classified based on their structural and functional characteristics?
3. How similar or different are the most frequent 3- to 4- lexical sequencies used by Turkish EFL learners from those found in English native speakers' (LOCNESS) corpora in terms of:
   a. structure,
   b. function, and
   c. diversity in their use?
4. What are the semantic domains of multi-word expressions in both TICLE and LOCNESS corpora? How similar and different are they from each other?

1.4. Significance of the Study

The current research provides many insights for both SLA researchers, English language teachers, instructors, material developers and publishers because lecical sequences are an integral part of oral and written language production and processing, and "provide a steppingstone into language development" according to Chenu & Jisa (2009, p. 27). Based on their findings, Biber et al. (1999) declare that "frequency should play a key role in the development of materials and in the choices that teachers make in language classrooms" (pp. 206–7). Warren and Greaves (2007) emphasize that the studies carried out on the use of multi-word units have fed and affected lexicography and the written performances of English language learners, and English language textbooks since

they shed light on the appropriate uses of these items, provide insights about the structures as well as functions they serve in the discourse and inform both learners, teachers, and publishers. From a psycholinguistic viewpoint, formulaic language has "a processing advantage over creatively generated language" for NNS and NS (Conklin & Schmitt, 2008, p. 72). Another study by Jiang and Nekrasova (2007) provides "prevailing evidence in support of the holistic nature of formula representation and processing in second language speakers" (p. 433). Likewise, Schmitt et al. (2004) also examined the psycholinguistic validity of corpus-derived recurrent clusters. They conclude that even though it is difficult to decide whether those clusters are stored in the mind as wholes by looking at the corpus data solely, we need both corpus and psycholinguistic approaches to explain language process and use (Schmitt et al., 2004).

In lexical bundle studies conducted by Biber and colleagues (Biber & Barbieri, 2007; Biber & Conrad, 1999; Biber, Conrad, & Cortes, 2003, 2004; Biber, Johansson, Leech, Conrad, & Finegan, 1999), it was revealed that conversation and academic prose differ in their patterns of lexical bundles. To exemplify, while most bundles in conversation are clausal, many bundles in academic prose are phrasal. Some studies of bundles generally compared bundles used in expert and non-expert writing. Some other studies focus on bundles used in the academic writing. More investigations of these structures in interlanguage corporas compiled from different L1 backgrounds are required to understand how these expressions are processed and used by EFL learners and to have insights and arrive at generalizations on the linguistic features.

Within the same line of reasoning, offering interpretations on how Turkish EFL learners produce multiword combinations in their written performances is of great importance and helps English language teachers understand the nature of these lexical bundles. Not only these structures reveal EFL learners' English proficiency levels but also provide the teachers with triggering input to create efficacious lessons and activities to support Turkish EFL learners' lexical competences. In addition, this research study examines not only structures and functions of the lexical bundles, but it also reveals semantic domains of these uses in the written performances of Turkish and American university students. Acquisition and production of these items also reveals the necessity of using corpora and authentic data both in and out of the classroom as recommended by Sinclair (1998) underlining that, "[c]orpora will clarify, give priorities, reduce exceptions and liberate the creative spirit" (p. 38) for the learner.

1.5. Operational Definitions and Key Terms

Applied Linguistics: "The use of knowledge about language to solve real-world problems" (Hunston, 2006, p. 245).

Computer Learner Corpus (CLC): "Electronic collection of authentic texts produced by foreign or second language learners" (McEnery & Wilson, 2001, p.177).

Concordance: "A comprehensive listing of a given item in a corpus (most often a word or phrase), also showing its immediate context" (McEnery & Wilson, 2001, p.197).

Corpus: "A corpus is an electronically stored collection of samples of naturally occurring language" (Hunston, 2006, p. 234).

Corpus Linguistics (CL): "A linguistics method which is founded on the use of electronic collections of naturally occurring texts, viz. corpora" (Granger, 2002, p.4).

Formulaic Sequences: Multiword units may also be referred to as formulaic language/sequences by some researchers (Conklin & Schmitt, 2008; Schmitt, 2004; Wray, 2000) or lexical bundles by some others (Biber, Conrad, & Cortes, 2004).

International Corpus of Learner English (ICLE): A learner corpus containing of argumentative essays written in English by learners from 16 different mother tongue backgrounds (Granger, Dagneaux & Meunier, 2009).

Learner Corpus (LC): A corpus containing written or spoken texts of second/foreign language (L2/FL) learners.

Lexical Bundles (LB): Lexical bundles are the most commonly co-occurring sequences of three or more words in a register that are neither fixed expressions nor idioms (Biber et al., 1999).

Log-Likelihood (LL) Analysis: It is a statistical analysis test that normalizes the sizes of the corpora compared and reveals the significances between frequencies.

Louvain Corpus of Native English Essays (LOCNESS): A native reference corpus containing British and American Students' written essays (Granger et al., 2009).

Phraseology: Phraseology may be defined as "the study of the structure, meaning and use of word combinations" by Cowie (1994, p. 3168).

The Turkish International Corpus of Learner English (TICLE): The sub corpus of ICLE, consisting of 200,000-words of argumentative writing by Turkish L1 learners of English as a Foreign Language (EFL) (Kilimci and Can, 2008).

Type-Token Ratio (T/T): T/t ratio is a helpful measure of lexical variety within

a written text or a speech (Williamson, 2013).

1.6. Limitations of the Study

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

CHAPTER II

REVIEW OF LITERATURE

2.1. Corpus Linguistics: An Overview

Corpus linguistics (CL) may be defined as a linguistic methodology based on the use of electronic collections of authentically occurring written and/or spoken texts, that is corpora, according to Granger (2002). She also stated that CL is not a new branch or theory of language, but it is an efficacious methodology of language research thanks to its authenticity (Granger, 2002).

Different researchers define CL in different ways such as Leech (1997) who considers CL more than a set of methods such as searching, retrieving, annotating, and analyzing electronic text data through computer tools such as taggers, parsers, and concordancers. For McEnery, Xiao and Tono (2006), it is a methodology, while Hoey (1997) views it as "the route into linguistics" (p. 6). Biber, Conrad and Reppen (1998) call it as "an approach" (p. 426), and similarly, Scott and Thompson (2001) define it as "a means for accessing resources" (p. 36).

As for Ludeling and Kytö (2009), CL has been described as a large taxonomic compilation of written and/or spoken language stored on a computer and used in linguistic analysis, and corpus data not only provides opportunities for linguistics research, but it also provides sources for research on teaching and learning of languages. In a similar vein, Bernardini (2004) highlighted that "more importantly perhaps, corpus has become less of a buzzword and more of a necessary, acknowledged reference source for students, linguists, language professionals (teachers, translators, technical writers, lexicographers etc.)" (p. 21).

When we consider the fields CL has influenced, these include culture, speech research, applied linguistics, sociolinguistics, grammar, semantics, lexical studies and more (McEnery & Wilson, 2001). For each area to be studied, a corpus may be designed with particular purposes such as the ones like historical, specialized, parallel, monitor, general, comparable, pedagogic and learner corpora for SLA and learning EFL/ESL (Hunston, 2002). Learner corpus is a collection of student production. Its aim is to identify in what respects learners differ from each other and from the native speakers.

2.2. Learner-Corpus Research

Learner corpus is defined as a systematic collection of machine-readable natural language data produced by L2 learners by Granger (2002, 2009). There are certain criteria such as learners' L1, sex, age, language background, and learning context to be considered in building the learner corpus (Dagneaux et al., 1998).

Since studies on learner corpora provide insights on "the needs of specific learner populations" (Meunier, 2002, p. 125) and help "to test teachers' intuitions about whether a particular phenomenon is difficult or not" (Granger, 2002, p. 22). To exemplify, Nesselhauf (2005) identified unidiomatic verb-noun collocations such as *drive a bike or reach an aim* found in German learner English corpus. This indicated that these types of combinations, *ride a bike* or *achieve an aim,* should be given particular attention while teaching and also in materials designed for German learners of English.

Learner corpus represents the learner language or "interlanguage" as coined by Selinker (1972). From interlanguage point of view, learner language is viewed as a non-linear, dynamic, and developing language system. As the linguists and researchers understand more about aspects of interlanguage, teachers and instructors help learners produce native-like language with more informed L2 instruction. In order to reveal the patterns of learners' interlanguage, various methods are employed on vast amounts of data by learner-corpus researchers (Granger, 1996). One of these kinds of methods is referred as Contrastive Interlanguage Analysis (CIA), (Granger, 1996), which compares learner language to other learner languages coming from different mother tongue backgrounds.

While some learner corpora studies try to evaluate the level of under- or overuse of particular linguistic features in learner language (Granger, 1996), some other studies try to uncover L1 transfer or interference. In fact, Contrastive Interlanguage Analysis (CIA) involves two types of comparison. Lorenz (1999) indicates that NS/NNS comparisons aim to shed light on non-native characteristics of learner writing with detailed comparisons of linguistic features in native and non-native speaker corpora. CIA also involves NNS/NNS comparisons. According to Lorenz (1999), "By comparing different learner populations, researchers improve their knowledge of interlanguage" (p. 12). Also, comparisons of learner data of different L1 backgrounds reveal interlanguage universals shared by various learner populations for researchers. These features may be developmental and L1 dependent. For example, in Granger and Tyson's (1996) study (of

connectors), overuse of sentence-initial connectors was stated to be developmental, and it was also emphasized to be an interlanguage characteristic of the three learner populations which are French, Dutch, and Chinese. The individual connectors having a wide variation of use among these learner groups provide evidence for interlingua characteristics.

One of the pioneering learner corpora is the International Corpus of Learner English (ICLE). ICLE comprises over three million words of essays written by advanced learners of English as a foreign language from various mother tongue backgrounds such as German, Spanish, French, Japanese, Polish, Chinese, Russian, Dutch, Swedish, Czech, and Finnish (Granger, 2003).

Another of the leading learner corpora in the field is the Cambridge Learner Corpus (CLC), which is a 52.5-million-word corpus of learner English compiled by Cambridge University Press (CUP) and Cambridge English Language Assessment since 1993 (Nicholls, 2003). CLC includes essays written in response to examination prompts by English language learners sitting examinations. CLC consists of over 200,000 exam scripts from students of 148 different mother tongues, living in 217 different countries or territories. Meta-information about the learner such as learner's L1, age, sex, education history and number of years the learner has studied English for are accommodated in each examination script.

Learner corpus has proven to be an invaluable source of data for many fields of SLA linguistic inquiry. For instance, Gilquin et al. (2007), Flowerdew (2001), O'Keeffe et al. (2007), Granger (2002) have highlighted the importance of using learner corpora to shed lights on difficulties NNS have while learning a language and as a reliable source of information that may be used in the classrooms both for teachers and learners.

Among the other major areas learner corpus research has informed, we could mention lexico-grammatical patterning, discourse and language features and non-native writing. In their study, for example, Milton (1997) also looked at the learners' use of fixed expressions and compared students' essays with a parallel native corpus. Finally, he found out students used few expressions repetitively in their writing.

Moreover, learner corpus has also helped to reveal some significant patterns in English for Academic Purposes (EAP) writing on the use of connectors (Flowerdew, 2001; Granger & Tyson, 1996). Other researchers such as Aijmer (2002) and Hinkel (2002) found out some features that reflected informal speech characteristic in learners' writing.

In another study focusing on the linguistic features of non-native writing, Cortes (2002a; 2004) presented the differences between student and professional writers in their use of lexical bundles specific to EAP, while Neff (2008) revealed pragmatic inappropriateness in non-native writing.

2.3. Corpora in English Language Teaching

O'Keeffe et al. (2007) posit that reviewing and understanding the process of second language development could be realized by analyzing learner corpora, which might bring light to the "long-held" notions of education and pedagogy bridging the gap between the cognitive linguistics and many other areas, such as sociolinguistics and translation. Recently, corpus-based analysis in language teaching provided empirical descriptions pertaining to language use, identification of the frequent patterns, the usage of particular forms and words in various registers (Biber & Reppen, 2002). Such findings from corpus-based studies offer invaluable insights and clarifications for both teachers and material developers (Biber & Reppen, 2002).

Using language corpora as a resource in language teaching has contributed to the acquisition of both implicit and explicit knowledge (Boulton, 2009). Ellis (2009) describes implicit knowledge as learning without paying attention. To him, it could be acquired from meaning-focused interaction. Whereas the explicit knowledge may be referred as learning with awareness through explicit instruction (Ellis, 2005) and "this explicit instruction can speed up implicit learning processes when supported with language items, patterns, and rules" (p. 162).

Corpus-based material development in teaching English as a Foreign Language (EFL) are considered, Aston (2000) emphasizes the effects of the studies on learner corpora on syllabus design as they specifically provide insights on "the needs of specific learner populations" (Meunier 2002, p. 125). One example of materials is *Longman Grammar of Spoken and Written English* (Biber et al., 1999) which has taken a corpus-based descriptive grammar approach by offering a corpus-derived perspective and information about frequencies. Likewise, many corpus-informed materials including course books or grammar books for both teachers and language learners have been written by the corpus linguists such as McCarthy, McCarten & Sandiford (2005) thanks to the multimillion-word corpora.

Also, Computer-Assisted Language Learning (CALL), which might include the

use of corpora in the language teaching context, is considered as a more direct application of CL in language teaching and applied linguistics. Learners can use a corpus themselves with corpus-based materials such as concordance lines or hands-on tasks as also exemplified in Johns' (1991) data-driven learning model.

Corpus may also benefit teaching translation. McEnery & Wilson (2001) claimed that using corpora in teaching translation had a positive effect on students' performances. The availability of multilingual parallel corpora facilitates translating by providing side-to-side examples of style and idiom in other languages and generating exercises that students can compare their own translations with an existing professional translation or original (McEnery & Wilson, 2001).

Another benefit of corpus studies in language teaching is teaching languages for specific purposes, which refers to the domain-specific teaching of a language for a particular area of application (Hyland, 2008a) such as teaching of English for Academic Purposes.

When pedagogical corpus applications are considered, a useful distinction should be made between direct and indirect applications according to Römer (2011) who asserts that corpora may inform us about what and when to teach. Corpora may influence language teaching in two ways: the direct and indirect approaches, as can be seen from Figure 1. While language learners and teachers' use of corpora themselves to uncover the patterns of language and words are referred as the direct approach (Bernardini, 2002), material designers and language teachers' use of corpora evidence provided by the researchers in designing courses or developing teaching materials is the indirect approach (Hunston, 2002).

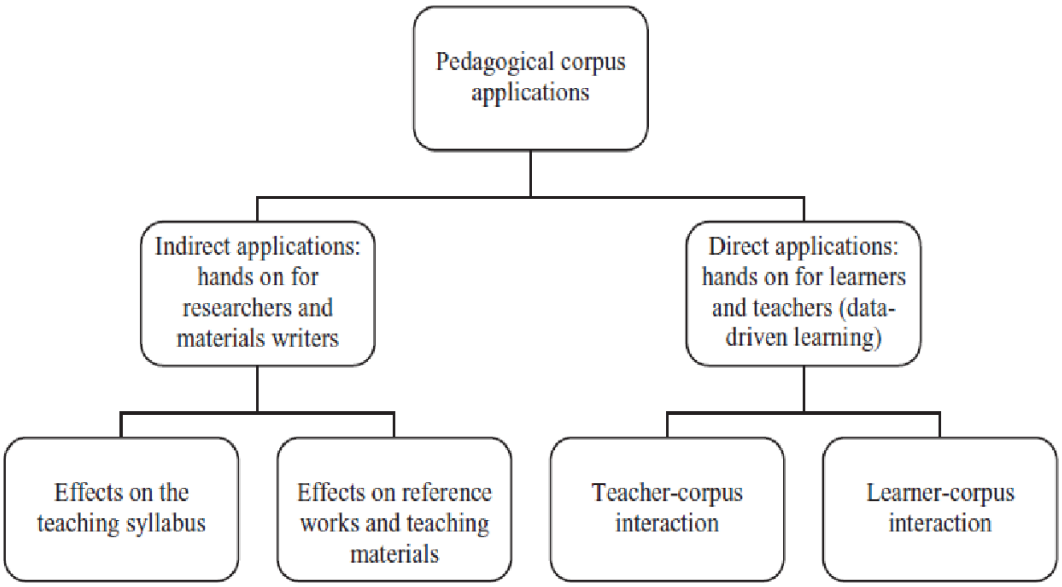


Figure 1. The use of corpora in second language learning and teaching
Source: Römer, 2011.

As can be seen from the figure, different direct and indirect applications may be identified. Learners and teachers can directly access corpora in the L2 classroom, and this "assists in the teaching process" (Fligelstone, 1993, p. 98). How something can be learned and taught is mostly affected by direct applications which involve both the learner and teacher while working with corpora or concordances (Römer, 2011). In contrast, indirect corpus applications inform syllabus design and design of teaching materials by the information provided by researchers.

2.3.1. Indirect Applications of Corpora in Language Pedagogy

The aim of the applied corpus research is to inform the teaching syllabus, highlight the importance of recurrent items, examine real-life language usages and compare the patterns in reference corpora with actual teaching materials such as coursebooks (Römer, 2011).

Corpora along with the evidence derived from it have the great potential to affect course design and the content of teaching materials (Römer, 2011). Existing pedagogical descriptions are revised by the "new evidence" according to Sinclair (2004c, p. 271), also, new decisions relating the selection of language phenomena, the advancement of the course, and the presentation of the items and structures are made (Mindt, 1981; Römer, 2005).

Designing the language syllabus and prioritizing teaching frequent items that learners are most likely to encounter are other significant effects of corpora. The first and the most important development in the corpus-informed ELT syllabus was the Collins COBUILD English Course (CCEC) (Willis & Willis, 1989), which was a COBUILD project in pedagogically oriented lexicography (Sinclair, 1987). The contents of this corpus-driven "lexical syllabus" are "the commonest words and phrases in English and their meanings" (Willis, 1990, p. 124). By focusing on the lexis and lexical patterns, the CCEC also revealed that language is highly patterned, comprising repeated word combinations and lexis and grammar are inseparable. COBUILD dictionaries, grammar books, and concordance samplers (Goodale, 1995; Sinclair, 1990, 1992, 2001) consist of real English and the needs of the language learner offering both teachers and learners far more reliable information about English language than any of the traditional reference grammars or other non-corpus-based dictionaries. What the COBUILD and other corpus-based reference works offer for learners (Biber & Conrad, 2009; Biber, Leech, & Conrad, 2002; Carter & McCarthy, 2006) is that they incorporate corpus-derived findings by taking frequency distribution and register variation into account and that they include authentic examples instead of invented ones. Another one is corpus-based Longman Grammar of Spoken and Written English (Biber et al., 2002) which provides learners with authentic language examples.

The related literature on using invented examples in teaching materials suggests that "invented examples can present a distorted version of typicality or an over-tidy picture of the system" (Kennedy, 1992, p. 366). Similarly, Sinclair (1991) noted that it is an "absurd notion that invented examples can actually represent the language better than real ones" (p. 5).

Even though materials and teachers lag behind the developments of the corpus studies, language learners may still decide to work with thoroughly corpus-based materials which offer improved representations of the language in question. There are also some attempts to produce textbooks based on corpus research and extracted from real-life data (Carter, Hughes & McCarthy, 2000; McCarthy, McCarten & Sandiford, 2005). Other developments include the world's first corpus-based English conversation program on television and accompanying published textbooks in Japan by Tono (2011).

Corpus studies also focused on analyzing language features which are problematic for learners. Some of these include *discourse particles* for German (Jones, 1997), *future time expressions* (Mindt, 1987; 1997) for English, *if-clauses* (Römer, 2004a), l*inking*

adverbials (Conrad, 2004) and *progressive verb forms* (Römer, 2005, 2007), and *reflexives* (Barlow, 1996). They compared naturally occurring English and German and the type of German or English in examined teaching materials. The results revealed all those mismatches in between and demanded corpus-informed points of view to the language-teaching syllabus.

Gavioli (2006) called for a need of using specialized corpora in English for specific purposes (ESP), noting that "working out basic items to be dealt with is a key teaching problem" (p. 23). Focusing on keywords in a specific discourse (e.g., medical research articles, newspapers) to be dealt within the courses may help solve this problem and assist teachers in "focus[ing] their efforts in terms of selection of language contents" (Pérez-Paredes, 2003, p. 1).

As for English for Academic Purposes (EAP), implementing Coxhead's (2000) Academic Word List (AWL), extracted from a corpus of academic writing, would be the most relevant and useful decision for EAP learners as it consists of the vocabulary items that are the most frequent and common in academic publications. In a similar vein, Simpson-Vlach and Ellis (2010) devised an Academic Formulas List (AFL), which includes word combinations occurring significantly more often in academic than in non-academic writing, by taking Coxhead's word list idea to phrase level. As both studies (AWL and AFL) suggest, EAP/ESP syllabus design may also be informed by specialized corpora.

2.3.2. Direct Applications of Corpora in Language Pedagogy

The direct approach focuses more on the teacher and learner (Römer, 2011). Learners and teachers themselves could try corpora and concordance tools to learn language patterns in an "autonomous" manner (Bernardini, 2002, p. 165) rather than relying on researchers or materials designers.

In terms of language teaching methodology, a more student-centered approach to learning and teaching and learners' becoming responsible for their own learning have been supported by corpus-based approaches. Direct corpus applications, especially in grammar and vocabulary classes, have been pioneered by Tim Johns in the 1980s. He advised to "confront the learner as directly as possible with the data, and to make the learner a linguistic researcher" (Johns, 2002, p. 108). The learners have also been referred as a "language detective" and originated "Every student a Sherlock Holmes!" maxim

(Johns, 1997, p. 101). This interaction between the learner and the corpora and between the teacher and the corpora is now known as data-driven learning (Johns, 1986; 1994).

Since then, some researchers have argued several ways in which corpora could be used by language learners and teachers. For instance, the significant effect of corpus-based learning using the BNC has been described by Bernardini (2002) who also defined corpora as "rich sources of autonomous learning activities of a serendipitous kind" (p. 165). Similarly, Kettemann (1995) also described the exploratory aspect of DDL by discovering concordance lines in an ELT classroom "motivating and highly experiential" for the learner (p. 30).

Some previous studies conducted using corpus-based activities, concordances or DDL in L2 learning context uncovered the effectiveness of these kinds of sources on learners' performance on vocabulary learning, mainly through collocation learning. Sun and Wang (2003) conducted a study on the effectiveness of inductive and deductive approaches on the learning of collocations by using concordance lines. They wanted to measure the relationship between the collocation patterns and learner performance by giving tests and making learners use concordance tools. They found out that language learners can be efficient discoverers of the target language with the help of concordances.

Another study by Chao (2010) analyzed the effectiveness of corpus-based activities on his learners' collocation learning. He investigated the effects of concordancing on the learning of receptive and productive collocations in an experimental study. Pre-tests and post-tests have revealed significant differences, the experimental group outperforming the control group in learning productive collocations.

Related with the effectiveness of direct corpus application, Sinclair (1997) noted that "[c]orpora will clarify, give priorities, reduce exceptions and liberate the creative spirit" (p. 38). Researchers such as Boulton (2009), Cobb (1997), and Granath (2009) indicated the effectiveness of DDL on the teaching of both grammar and vocabulary noting that DDL encourages learners to explore the language for themselves, raises the learners' language awareness and has great pedagogic potential.

In his study of effectiveness of using corpora in L2 academic writing, Yoon (2008) claimed that "students assumed more responsibility for their writing and became more independent writers" (p. 31). Likewise, Johns et al. (2008) noted that "CALL activities [including DDL] did have a positive impact on students' English learning" (p. 503). Also, in his study, Boulton (2009) stated that "corpus samples led to more successful results than traditional pedagogical resources" such as bilingual dictionaries or textbooks (p. 50).

Focusing on concordancing was a useful way “to mimic the effects of natural contextual learning” (Cobb, 1997, p. 314). This also proved to be a useful source both in teaching translation teaching according to Bernardini (2004b) and error correction in second language writing (Bernardini, 2004a). All these descriptive and empirical studies reveal corpora may provide information that grammar book or a dictionary may not provide (Römer, 2011).

2.4. Phraseology

2.4.1. Phraseology and Linguistic Theory

Hunston (2002) who is the first corpus linguist to discuss “some examples” which she later categorized as “[c]ollocation”, [p]hrases and variation”, “[t]he tendency of certain verbs to occur in the passive rather than the active, or in the negative rather than the positive” (i.e., what is usually referred to as colligation), and “[t]he occurrence of complementation patterns” (p. 137f.). Even though Gries (2008) agrees with the definitions provided, he also adds other six parameters to define phraseology as can be seen from the table below:

Table 1.

Six Parameters of Phraseology

i. the nature of the elements involved in a phraseologism;	the co-occurrence of a form or a lemma of a lexical item and any other kind of linguistic element, which can be, – another (form of a) lexical item (*kith and kin* is a very frequently cited example of a nearly deterministic co-occurrence of two lexical items, as is *strong tea*); – a grammatical pattern (as opposed to, say, a grammatical relation), i.e., when a particular lexical item tends to occur in/co-occur with a particular grammatical construction.
ii. the number of elements involved in a phraseologism;	Phraseologisms can consist of only two elements (such as word pairs) or can include a larger number of elements.
iii. the number of times an expression must be observed before it counts as a phraseologism;	Observed frequencies must exceed frequencies expected on the basis of chance (significantly or just at all). An expression can be considered as phraseologism only if its observed frequency of occurrence is larger than its expected one.
iv. the permissible distance between the elements involved in a phraseologism;	N-gram-based studies in natural language processing concerns itself only with immediately adjacent elements, however, a more widespread broader perspective which also recognizes discontinuous phraseologisms should be adopted.
v. the degree of lexical and syntactic flexibility of the elements involved;	Studies that are only concerned with completely inflexible patterns such as *by and large* can be distinguished from studies that include relatively flexible patterns such as *kick the bucket* (which allows different tenses but no passivization), studies (also) involving partially lexically filled patterns such as the *into*-causative ([VP V DO *into* V-*ing*]), and studies including completely lexically unspecified and thus maximally flexible expressions, such as the English ditransitive pattern [VP VOBJ1 OBJ2]. Only the last of these should be excluded because they do not involve at least one lexically specified element (as required by the first parameter).
vi. the role that semantic unity and semantic non-compositionality / non-predictability play in the definition	The elements of a phraseologism are usually assumed to function as a semantic unit, i.e., to have a sense just like a single morpheme or word. While semantic unity is required, non-compositional semantics is not required to be counted as phraseologism.

Adapted from Gries (2008, p. 4-6)

According to this, a phraseologism can be defined as:

> The co-occurrence of a form or a lemma of a lexical item and one or more additional linguistic elements of various kinds which functions as one semantic unit in a clause or sentence and whose frequency of co-occurrence is larger than expected on the basis of chance (Gries, 2008, p. 6).

Moreover, as the first parameter states, the lexical items and grammatical patterns are not distinguished from each other. Sinclair (2008) also states that grammar and lexis/semantics are not divided; "it takes a holistic view, starting with phrases taken directly from text; since there is not a strong tradition of the description of this kind, it allows, even encourages, a wide range of approaches, as the papers which follow testify" (p. xv). However, phraseology does not take place in the traditional language context according to Sinclair (2008).

Highly frequent co-occurring expressions like *of the* or *in the* are not included in the definition since these expressions do not function as a semantic unit (Gries, 2008). For instance, it can be posited that *in spite* is a phraseologism if semantic unity were not required as it contains two words (number and nature of elements) that co-occur more often than expected by chance. However, for Gries (2008), statistically more sophisticated approaches would recognize *in spite of* as a more reasonable assumption in the 'real' phraseologism.

2.4.2. Phraseology in Corpus Linguistics

According to Gries (2008), some most central notions such as *word clusters*/n-*grams* and *collocations* in corpus linguistics could be compared to phraseologisms depending on the six criteria explained above. *Word clusters*/n-*grams* and *collocations* refer to frequent co-occurrences:

– nature of the elements: words;
– number of elements: *n* (usually, that means 'two or more');
– frequency of occurrence: sufficiently frequent to be recognized as a combined element;
– distance of elements: for clusters/*n*-grams, the distance is usually 0 (i.e., the elements are immediately adjacent); for collocations, the distance between the

> elements involved can vary, but usually exhibits one or a few preferred distances;
> – flexibility of the elements: for clusters/*n*-grams, there is usually no flexibility; for
> collocations, one usually allows for some flexibility: the collocation of *strong* and *tea* would be instantiated both by *strong tea* and *the tea is strong;*
> – semantics: *n*-grams are usually retrieved for natural language processing purposes
> where the issue of non-compositional semantics is only sometimes relevant; for collocations, researchers differ as to whether they require some non-predictable behavior (*strong tea* is acceptable, but *powerful tea* is not) or not. (Gries, 2008, p. 16)

Another notion in corpus linguistics is the *pattern* that involves additional parameters of the above set, viz. the parameter of non-compositionality/non-predictability. Among many, one of the major applications of corpus-driven research in phraseology is the 'pattern grammar' framework, referred to as a corpus-driven approach to the lexical grammar of English (Hunston & Francis 2000). Hunston & Francis (2000) define a pattern as:

> The patterns of a word can be defined as all the words and structures which are regularly associated with the word and contribute to its meaning. A pattern can be identified if a combination of words occurs relatively frequently, if it is dependent on a particular word choice, and if there is a clear meaning associated with it (p. 37).

As can be inferred**,** a phraseologism is characterized by a sufficiently high frequency of co-occurrence (Hunston & Francis, 2000). A corpus-based approach to phraseologisms and phenomena,including c*ollocations, colligations,* and *patterns* simply provide with the required frequency data.

As for the identifications of phraseologisms in corpus linguistics, Stubbs and his colleagues (2001; 2002), generated a frequency list of *n*-grams, which are uninterrupted sequences of word forms with the upper limit of *n* usually being five. In some other studies, researchers also extracted concordances of one or several word forms and interpreted the data for generalizations about the preferred uses of the word forms (Sinclair, 1991; Hunston & Francis, 2000).

2.4.3. Phraseology and Lexicography

Phraseology was defined as "the study of the structure, meaning and use of word combinations" by Cowie (1994, p. 3168). Linguists distinguish phraseological units from non-phraseological ones by functions as word combinations are in various shapes and forms. For Granger and Meunier (2008), being represented as a subfield of lexicology, phraseology deals with word combinations, their semantic non-compositionality, institutionalization, lexical restrictions, and syntactic fixedness which determine the classifications of these multi-word units (MWUs). Phraseology takes place both in different theoretical frameworks within linguistics and in the methodological paradigm of corpus linguistics (Gries, 2008). Corpus linguistics has led lexis-grammar research to study the meaning beyond the word by bringing the phraseology to the front lines in language analysis. Some of the most influential research in the field are discussed below.

One of the earliest milestone studies belongs to Pawley and Syder (1983) who claimed that formulaic language is the key to nativelike proficiency and second language fluency. They further noted that language is acquired in chunks and word strings are stored and retrieved as wholes from memory to fit the functions and meanings.

Within the same line of reasoning, John Sinclair has paved the way to corpus-driven lexico-grammatical research with two concepts: the idiom principle and lexis grammar. According to the idiom principle, phraseological tendency of language in which words do not appear in isolation and make connections with one another to create meanings is discussed (Sinclair, 2004). The idiom principle differs from the open-choice principle in which words are individually chosen to fill certain slots in the sentence. For Sinclair (1991), "a language user has available to him or her numerous semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable into segments" (p. 110). Texts are made up of multiword expressions which establish single choices in the mental lexicon.

After many studies in the field proved that lexis and grammar are inseparable, Sinclair (2004) proposed the notion of *lexical grammar* which he has defined as "an attempt to build together a grammar and lexis on an equal basis" (p. 164). According to this notion, structure and meaning are counted as one.

Another milestone in the field is Hunston and Francis' pattern grammar following Sinclair's concept. Pattern grammar has been developed from a study of 250-million-word Bank of English claiming two fundemental points about the grammar of words or

patterns: “firstly, that all words can be described in terms of patterns; secondly, that words which share patterns also share meanings” (Hunston et al., 1997, p. 209). The first statement may be exemplified by patterns like V and Vn for the verb *to eat*, or the ones associated with the impersonal *it.* The second one may be exemplified by patterns such as V *by* -ing, which the V slot is usually filled with verbs falling into one of two meaning groups: “to start” or “to end” (begin by saying) and “to respond to or compensate for something” (*atone by fasting*) (Hunston et al., 1997, p. 213). While some words have several patterns for the same meaning, others have particular patterns for particular senses. In addition, some others may have many meanings which could be clarified using the distinct patterns they occur in.

In Sinclairian contextualist tradition, Hoey (2005) proposed the notion *of lexical priming* in which the traditional roles of vocabulary and syntax are reversed. Grammar is an output of complex lexical structure rather than constraining lexis. The notion of ‘collocation’, the property of language whereby two or more words seem to appear frequently in each other’s company (e.g., ‘inevitable’ and ‘consequence’), gives a clue about how language is organized. Using concrete statistical evidence from a corpus of newspaper English, travel writing and literary texts, Hoey (2005) argues words are ‘primed’ for use through our experience with them, so “…everything we know about a word is a product of our encounters with it. This knowledge explains how speakers of a language succeed in being fluent, creative, and natural” (Hoey, 2005, p. i).

Within the theory of Lexical Priming, Hoey (2004) also emphasizes that “the only explanation that seems to account for the existence of collocation is that each lexical item is primed for collocational use” (p. 386). Priming is not simply between words, multi-word expressions, for instance, it may also have collocations that are distinct from those of the words that make them up. Sinclair (2004) exemplifies with his well-known analysis of the expression naked eye to illustrate how phrases and collocations can combine in texts. In this context, the words within the multi-word expression are primed to co-occur, and the co-occurring combination is also primed for particular collocational patterns. Hoey (2005) labels *nesting* to identify this phenomenon, simplifying the memory’s task. Primings nest and combine. For example, *winter* collocates with *in*, producing the phrase *in winter* (Hoey, 2005), however, this phrase has its own collocations, which are separate from those of its components. Therefore, in winter collocates with several forms of the word BE (i.e., is, was, are, were, etc.).

2.4.4. Phraseology and English Language Teaching

Benefiting from the promising potential of corpora, Collins COBUILD English Language Dictionary (Sinclair, 1987) has changed dictionary-making and pioneered the use of language corpora as a standard tool for lexicographers (O'Keeffe, McCarthy & Carter, 2007). Lexicographers could have access to huge amounts of data by using corpora. They could also explore the occurrences of expressions and their frequencies of usage (Svensen, 2009).

Williams (2003) states that lexicographers rely on the patterns emerging from corpus in corpus-driven lexicography. Thanks to the developments in this particular field of study, more contextual dictionaries have been written to cover the definitions of both lexical and grammatical words based on their actual language uses. Furthermore, developments of more efficient ways of storing, cross-referencing source materials, transferring and accessing have been brought to life by lexicography (Williams, 2003).

When exposed to a word, the mind subconsciously keeps a record of the context and co-text of the word. As we re-encounter the word (or syllable or combination of words), we build up a record of its collocations. We are primed by each encounter, therefore; while using the word (or syllable or combination of words), the contexts we previously encountered are replicated (Hoey, 2005). With the theory of priming, Hoey (2004) also explains "the only explanation that seems to account for the existence of collocation is that each lexical item is primed for collocational use" (p. 386). Collocation is not only explained for lexical priming, but it also describes everything we know about a word. As the mind has a mental concordance of every word which it encounters, a concordance is richly annotated for social, generic, physical, discoursal and interpersonal context (Hoey, 2005). According to him, these are all connected to word association games.

2.5. Categories of Word Combinations

As shown in Figure 2, a main subdivision may be seen between two main extraction methods: n-gram analysis and co-occurrence analysis (Stubbs, 2002).

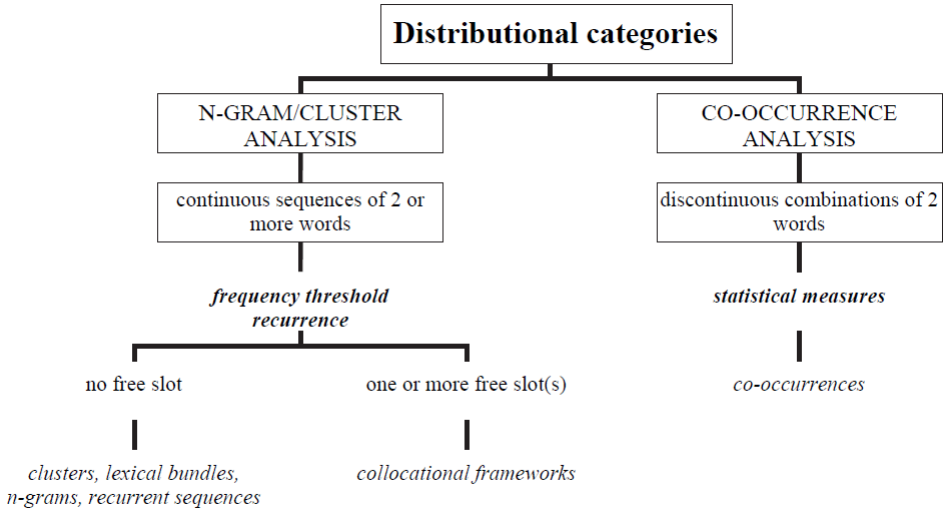


Figure 2. Distributional categories of word combinations
Source: Stubbs, 2002, p. 39

According to this, n-gram analysis allows for the extraction of continuous recurrent sequences of two or more words, "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" as Biber et al. (1999, p. 990) define. The extracted sequences are called n-grams (Stubbs 2007a, 2007b), lexical bundles (Biber & Conrad 1999; Biber et al. 2003; Biber 2004), clusters (Scott and Tribble 2006), chains (Stubbs, 2002), or recurrent word combinations (Altenberg, 1998). Some examples of retrieved sequences are "*I don't know what, I thought that was, can I have a, in the case of, on the other hand, the use of, the fact that"* (Biber et al., 1999, p. 1002-3).

2.6. Formulaic Language: Definition, Categories, and Identification

A consensus on the definition of formulaic language clarifies that the items have to be:

1. multi-word
2. have a single meaning or function
3. be prefabricated or stored and retrieved mentally as if a single word (Wood, 2015, p. 3)

Pawley and Syder (1983) clearly stated that formulaic language is likely to be a key indicator of second/foreign language fluency and native-like performance. Even though we have many creative ways of producing utterances in the language such as

saying, “what is the nature of your current well-being?” instead of “how are you?”, we prefer language in “chunks” due to processing restrictions and acquire language by storing and retrieving word strings as wholes from long term memory according to Pawley and Syder (1983). Later in 1991, Sinclair stated the *idiom* and the *open choice* principle, as can be seen in p. 22, that “most texts are largely composed of multiword expressions that constitute single choices in the mental lexicon” (Sinclair, 1991, p. 73).

Several terms have been used to refer to formulaic language such as formulas, formulaic sequences, lexical bundles, chunks, formulaic speech, conventionalized forms, multiword units, prefabricated routines, and ready-made utterances (Schmitt, 2004; Wray, 2002). A psycholinguistic definition for formulaic sequences is a sequence of words that appears to be prefabricated, “that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar” (Wray, 2002, p. 9). According to this definition, formulaic sequences are retrieved from the memory as a single word in the form of a chunk. This study uses the terms interchangeably, which can be simply defined as multiword units of language. In addition, Wray and Perkins (2000) show that formulaic language can be labeled as many as forty terms as can be seen from the Table 2 below.

Table 2.

Terms Used to Refer to Formulaic Language

Amalgams	Noncomputational
Automatic	Nonproductive
Chunks	Petrification
Cliches	Praxons
Co-ordinate constructions	Preassembled speech
Collocations	Prefabricated routines and patterns
Composites	Ready-made expressions
Conventionalized forms	Ready-made utterances
FEIs	Rote
Fixed expressions	Routine formulae
Formulaic language	Schemata
Formulaic speech	Semi-preconstructed phrases that constitute single choices
Formulas/formulae	Sentence-builders
Fossilized forms	Stable and familiar expressions with specialized subsenses
Frozen phrases	Synthetic
Gambits	Unanalyzed chunks of speech
Gestalt	Wray and Perkins (2000, p. 3)
Holistic	
Holophrases	
Idiomatic	
Idioms	
Irregular	
Lexical(ized) phrases	
Lexicalized sentence stems	
Multiword units	
Noncompositional	

As can be seen from the table, formulaic language refers to many concepts such as collocations, phrasal verbs, idioms, lexical bundles, n-grams, conc-grams and lexical phrases (Wray and Perkins, 2000). Since there are many categories under the formulaic language term, some categories overlap with others and classifications are arbitrary (Wood, 2015).

The studies in the field provided two perspectives for identifying the formulaic language in terms of research methodology:

- The frequency-based approach (Firth, 1951; 1957) which is concerned with the statistical likelihood of words appearing together.
- The phraseological approach (rooted in Soviet phraseology) leans more on restrictive descriptions of multiword units, with a narrower view of what to label as a collocation.

Formulaic sequences make up a large portion of discourse (Schmitt & Carter, 2004). Erman and Warren (2000) reported that different types of word combinations made up 58.6% of the spoken corpus analyzed in their study and 52.3% of the written corpus.

Apart from having a high frequency, the use of these recurring word combinations has been considered a sign of proficient language use of particular registers. Wray (2002) also stated that formulaic sequences let language users relate their identity with a group, and the use of these expressions diminishes the cognitive load in reading or listening process. Haswell (1991) stated that the use of recurrent fixed expressions reflects certain maturity in writing, while the lack of these expressions is a marker of novice writers. Cortes (2004) underlined that the frequent use of fixed expressions "seems to signal competent language use within a register to the point that learning conventions of register use may in part consist in learning how to use certain fixed phrases" (p. 398).

2.6.1. Acquisition of Formulaic Language

Many corpus-based research studies revealed that non-native academic writers have difficulty in acquiring native-like lexical bundles. Non-native and native writing display significant variations in terms of lexical bundles. According to recent studies, NNS overuse or underuse certain lexical bundles in their written performances, and they use more limited and less varied items compared to native speakers (Cortes, 2004; Chen & Baker, 2010; Allen, 2009). Furthermore, the target bundles produced by students does

not correspond to the bundles used by professional writers (Hyland, 2008b; Adel & Erman, 2012; Öztürk, 2014).

One of the pioneering studies on the acquisition of formulaic language, Yorio (1980) studied formulaic language in adult second language development. The study revealed that adult learners' use of formulaic language was limited. They only used those items as a written performance strategy and to save effort and attention.

Schmidt (1983) carried a study on English language development of a Japanese adult in Hawaii. In his well-known study, Schmidt found out that formulaic sequences were an indispensable aspect of language acquisition. As a communication strategy, the participant had used several number and range of formulaic sequences. It was also reported that there was resistance from the subject regarding the error correction. However, the participant succeeded in building his second system through using formulaic sequences.

In another study by Ellis (1996), it was reported that short-term repetition, rehearsal and memorized sequences constitute much of language acquisition process. Long-term storage of language sequences enables a learner to use them for semantic references and access more automatically. This allows for more fluent language production, as attention is given to dealing with conceptualizing and meaning.

In a similar vein, Bolander (1989) conducted a study on learners of Swedish as a second language who frequently used prefabricated language units containing target language structures. He found that formulaic sequences contributed to the ease of learning and use.

Studies in the related literature suggest that adults in naturalistic L2 learning environments, like children, acquire and use formulaic sequences. However, knowledge of L1, diverse acquisition contexts, the established cognitive and learning styles of adults and other factors also have an influence particularly on the use of formulaic sequences. While some adults may think more analytic and attempt to infer rules from chunked units or input, others, such as Schmidt's (1983) L2 learner, may heavily rely on acquired formulas and not attempt to break them down to analyze. Degree of literacy and type and instruction may also play a part.

Studies strongly imply that lexical bundles are acquired incrementally just like single words as Schmitt (2000) and Nation (2001) argue. Learners need a great deal of repeated and systematic exposures to acquire these bundles (Schmitt et al., 2004; Li & Schmitt, 2009). To help learners enhance their retention and acquisition of these multi-

word items, noticing, retrieval and generative activities including substitution tasks (Salazar, 2014), rephrasing (Peters & Pauwels, 2015), or writing activities (Nation, 2001) may be used in effective ways.

2.7. Lexical Bundles Research: Definition and Identification

Lexical bundles are an important type of formulaic sequences. The term 'lexical bundle' was first created by Biber et al. (1999) in the thirteenth chapter of the Longman Grammar of Spoken and Written English (LGSWE). Biber et al. (1999) describe lexical bundles as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" and as "simply sequences of word forms that commonly go together in natural discourse" (p. 990). Cortes (2004) also gives another consistent definition of lexical bundles as "extended collocations of three or more words that statistically co-occur in a register" (p. 400). Biber & Conrad (1999) identify lexical bundles as "the most frequent recurring lexical sequences…, which can be extended collocations: sequences of three or more words that show a statistical tendency co-occur" (p. 183). Additionally, Biber et al. (1999) note that "lexical bundles can be lexical building blocks that tend to be used frequently by different speakers in different situations" (p. 991).

Some examples of lexical bundles are kinds of expressions such as *as a result of, take a look at, on the other hand,* and *I don't know, to be able to, are you going to, I don't know if, at the same time, have a lot of, you know it was*. These items can be identified through computer software which run through groups of three or more words that co-occur frequently in the texts. When compared to other types of occurring word combinations, lexical bundles are identified empirically rather than perceptually in a corpus, and they are the most frequent multi-word sequences in a corpus with particular functions in discourse.

While identifying these items, the frequency-based approach is a foundation as mentioned above (p. 28). There are three major components of lexical bundle research:

1. Identification of these items are made by corpus analysis tools such as WordSmith Tools (Scott, 2007) by using the criterion of frequency cut-off.
2. A distribution or range criterion is set besides the frequency; an item must appear in at least five of the texts in the corpus to be counted as a lexical bundle.
3. Assigning functions to the sequences identified by frequency and range

criteria; sequences are categorized according to functions in the discourse (Biber et al., 1999).

Altenberg (1993, 1998) is one of the first researchers who conducted lexical bundles research by using frequency as the primary criteria in identifying particular word combinations, and by categorizing them functionally. The actual frequency cut-offs in lexical bundle research vary from ten occurrences (Biber et al., 1999; Biber, 2006), to forty occurrences per million words (Biber, Conrad, & Cortes, 2004). Lower cut-offs may be used in smaller corpora (Biber et al., 1999) and determining the frequency cut-offs is based on the earlier researchers rather than statistical or empirical set of standards. Also, the range and length criteria are two pivotal components of these multiword items (MWI henceforth). For a sequence to be labeled as lexical bundle, it has to appear in at least three to five texts (Biber & Barbieri, 2007). Some other researchers such as Hyland (2008) employed a percentage criterion which required a sequence to appear in at least 10 percent of the texts in a corpus.

Even though lexical bundles are defined within three parameters which are frequency, range and function, some researchers use additions or modifications to these parameters while studying multiword sequences (Wood, 2015). This entails the use of other terminologies. For example, Simpson-Vlach and Ellis (2010) referred their units as "formulaic language"; Wood and Appel (2014) and Liu (2012) used the term "multiword constructions." In addition, these combinations "are so common, it might be assumed that lexical bundles are simple expressions, and that they will therefore be acquired easily" (Biber and Conrad, 1999, p.188). These constructions are mostly used in published academic writing, including journal articles. Also, specific types of bundles are characteristic of particular disciplines (Cortes, Jones, & Stoller, 2002). Biber and Conrad (1999) state that acquisition and use of lexical bundles do not come naturally since they require focused instruction.

Lexical bundles are quite common, forming an important part of discourse. According to Biber et al. (1999), 21% of all the words in their academic prose corpus are recurrent lexical bundles. Besides their recurrent nature, lexical bundles also have particular characteristics which distinguish them from other types of multi-word expressions such as *collocations* and *idioms*. One of these characteristics is that "most lexical bundles are not idiomatic in meaning and not perceptually salient" (Biber & Barbieri, 2007, p. 269). Another characteristic is that lexical bundles are not usually

complete structural units like *in the case of* and *the base of the* (Biber & Barbieri, 2007), but they are parts of longer structures. Both function words and content words are found in lexical bundles, and it is clear in the literature that they vary across different disciplines (Hyland, 2012).

2.7.1. Structural Classifications of Lexical Bundles

Lexical bundles have strong grammatical correlates even though they are not complete structural units (Biber et al., 1999). Biber, Conrad, & Cortes (2004) has developed a categorization of lexical bundles according to their structures and function the discourse. This current study will use this structural and functional taxonomy developed by Biber, Conrad, & Cortes (2004, p. 381-396) which are summarized in Tables 3 and 4.

Table 3.

Structural Classification of Lexical Bundles

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

Adapted from Biber et al., 2004, p. 381-396

As can be seen from the table, bundles have been classified according to three major structural features. The first structural type incorporates verb phrase (VP henceforth) fragments. These types of bundles begin with a subject pronoun, and then followed by a verb phrase, as in *I'm not going to, that's one of the, it's going to be,* and

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

The second major structural type resembles the first category in how it incorporates verb phrase elements, however, these types of bundles also incorporate dependent clause fragments. To exemplify, *I want you to, what I want to,* and *if we look at.*

As for the last major structural type, only phrasal components are included. Most of these bundles comprise noun phrase (NP henceforth) components ending with the start of a post modifier as in *the end of the, the way in which, those of you who* and *a little bit about.* Prepositional phrase (PP henceforth) components with embedded modifiers comprise the rest of the bundles of this type. The examples are *of the things that, as well as the, at the end of.*

2.7.2. Functional Classifications of Lexical Bundles

Biber et al. (2004, p. 389-396) identified three primary discourse functions for lexical bundles in English: 1) stance expressions, 2) discourse organizers, and 3) referential expressions. Bundles that serve similar functions have been grouped together depending on their uses and meanings which can be seen below.

Table 4.

Functional Classification of Lexical Bundles

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

Adapted from Biber et al., 2004, p. 381-396

Three categories within the functional taxonomy are presented in detail below.

2.7.2.1. Stance Bundles

Stance bundles express a frame for attitude or assessment of certainty, which may be categorized in two*; epistemic* or *attitudinal / modality* (Biber et al., 1999). Epistemic stance bundles express the knowledge state of the information that follows, while attitudinal/modality stance bundles express writers' or speakers' attitude toward the content of what follows. Stance bundles may also be classified as personal or impersonal according to the judgments and attitudes. While personal stance bundles refer to the speaker/writer with the subject 'I' as in *I don't know what*, impersonal stance bundles resemble to personal stance bundles. However, they do not refer the speaker/writer directly, such as *it is possible to* and *can be used to*.

Majority of them being personal, epistemic bundles may convey uncertainty such as *I don't think, I don't know what;* or a sense of imprecision like *stuff like that* and *what do you think.* As for the impersonal epistemic bundles, degrees of certainty are expressed such as *in the fact that* and *are more likely to be*.

Also, attitudinal/modality stance bundles are mostly personal, which express speakers' attitudes towards actions. Examples of this type are *I am not going to* and *I want you.* These also have been divided into subcategories: desire, obligation/directive, intention/prediction, and ability. Desire bundles frame self-motivated desires and wishes or ask about other participants' desires. Examples include *I would like to, what I want to do, I don't want to, what I want to do.*

Obligation and directive bundles have a second-person pronoun. They express obligations and directions. *I want you to, you need to know, you might want to, you have to do,* and *take a look at* are some bundles in this group.

Being overtly personal, intention/prediction bundles indicate the speaker/writer's intention to do something. Some examples may be *what we're going to, is going to* (the impersonal bundle).

Lastly, ability bundles express the ability to do or accomplish something, such as *to come up with* or *to be able to*. Classroom discourse includes most of this type of bundles (Biber et al., 1999).

2.7.2.2. Discourse Organizers

This type of bundles organizes the flow of ideas like relationships between prior and coming discourse. They help to introduce topics, elaborate, and clarify topics. Examples include *I want to talk about, if you look at* or *going to talk about.* These types of bundles have two major functions: topic introduction/focus and topic elaboration/clarification (Biber et al., 1999).

Topic introduction/focus bundles introduce a new topic. They may also express intention and desire. For example, *I want to talk about, if we look at, what I want to do* are the bundles in this group.

As for topic elaboration/clarification bundles, additional explanation or clarification can be made with these types of bundles, as in *you know, I mean.* These bundles may also compare and contrast, such as *as well as the* and *on the other hand.*

Identification/focus bundles mostly have discourse organizing function even though Biber, Conrad and Biber (2004) categorize them as referential expressions. Examples are *that's one of the, those of you who, one of the things.*

2.7.2.3. Referential Bundles

Referential bundles comprise a large number and range of lexical bundles. They usually refer to the textual context, physical or abstract entities as in *the nature of the* and *that's one of the.* This type has four sub-categories: *identification/focus, imprecision, specification of attributes,* and *time/place/text reference* (Biber et al., 2004). All of them serve different purposes.

The first type identification/focus bundles mostly characterize the subsequent proposition as important and examples are *and this is a, one of the things and that's one of the.* The second type of these bundles which are imprecision bundles signal imprecise reference or there may be similar references which may be provided as in *or something like that* and *things like that.*

Another subcategory of referential bundles identifies specific qualities of the noun that follow. While some of them signal quantity or amount including *have a lot of, a little bit of, a little bit about;* others describe the size and form of the head noun like *in the form of* and *the size of the;* some others describe abstract characteristics as in the examples *in terms of, on the basis of, the nature of the.* Lastly, time/place/text-deixis bundles refer to particular time, place and/or location in the text. Examples include *in the United States, the end of the year…*etc.

2.7.3. Major Studies on Lexical Bundles

Lexical bundles and their use have been investigated in many areas including classroom discourse and textbooks, multidisciplinary texts and published papers, discipline-specific articles, and student essays. Studies focused on lexical bundles in different registers, as in academic prose and conversation (Biber et al., 1999; Biber, Conrad & Cortes, 2004), different disciplines like biology and history (Cortes, 2004), the usage of bundles between native and non-native writings (Ädel & Erman, 2012; Öztürk, 2014; Muşlu, 2015) or on pedagogical aspect of these expressions (Jones & Haywood, 2004; Li & Schmitt, 2009). Table 5 lists major corpus-based studies along with their focus of the studies and corpora used.

Table 5.

Major Corpus-based Studies on Lexical Bundles

Researcher(s)	**Year**	**Focus of the Study**	**Corpus**
Biber, Johansson, Leech,	1999	Lexical bundles in two registers: conversation and academic prose	LSWE Corpus
Conrad, & Finegan Cortes	2004	Lexical bundles in published and student disciplinary writing at three levels	Professional and student writings
Biber, Conrad & Cortes	2004	Lexical bundles in classroom teaching and	T2K-SWAL Corpus textbooks
Biber & Barbieri	2007	Lexical bundles in university spoken and written registers	T2K-SWAL and LSWE
Hyland	2008a	Lexical bundles and disciplinary variation	Research articles, doctoral dissertations and MA theses
Hyland	2008b	Academic clusters in published and postgraduate writing	Research articles, doctoral dissertations and MA theses
Allen	2009	Lexical bundles in learning written discourse	ALESS Learner Corpus
Byrd & Coxhead	2010	Creating a highly important list of lexical bundles for teachers	414 academic texts in four disciplines
Chen & Baker	2010	Lexical bundles in L1 and L2 academic writing	FLOB and BAWE
Simpson-Vlach & Ellis	2010	Pedagogically useful list of formulaic sequences for academic speech and writing	MICASE, BNC and Hyland's (2004) research article corpus
Wei & Lei	2011	Lexical bundles in academic writing of advanced Chinese EFL learners	Doctoral dissertations and published journal articles
Adel & Erman	2012	Lexical bundles in academic writing by native and non-native speakers	SUSEC
Martinez & Schmitt	2012	A Phrasal Expression List	BNC
Chen & Baker	2016	Lexical bundles in rated learner essays, CEFR B1, B2 and C1	Longman Learner Corpus (LLC)
Shin	2018	Lexical Bundles in Argumentative Essays	L1-Korean-speaking EFL learners' essays

As can be seen from the table, the first study was conducted by Biber et al. (1999) who identified and described the patterns of occurrences in Longman Spoken and Written English Corpus (LSWE), which is a corpus of American and British English conversation and academic prose. They compared and contrasted the conversation and academic

genres. They concluded that there were almost ten times as many 3-word bundles as 4-word bundles. However, 5 and 6-word bundles were much less common. Also, there were many more lexical bundles in conversation prose than in academic writing according to their findings.

The second major study investigating the use of bundles in university classroom teaching and textbooks was conducted by Biber et al. (2004). They used T2K-SWAL Corpus, which is the TOEFL 2000 spoken and written academic language corpus in their study. The corpus includes both spoken and written activities associated with academic life, classroom teaching, textbooks, study groups, university catalogues sampled from six major academic disciplines which are business, engineering, humanities, social and natural sciences, and education. Findings of the study state that classroom teaching use twice as many LBs than conversation and four times as many LBs as textbooks. LBs are much more common in both conversation and classroom teaching than in the written registers. In this study, they also revised and extended the structural and functional classifications of the bundles.

A later study by Biber & Barbieri (2007) expanded the scope of registers in Biber and his colleagues' research (1999, 2004) and examined nine lexical bundles in spoken and written registers in total. They stated that lexical bundles differ from other lexico-grammatical linguistic structures regarding physical mode (spoken/ written differences) and communicative purposes (Biber & Barbieri, 2007). Physical mode affects the grammatical features while mode and communicative purposes affect the use of lexical bundles. Lexical bundles are "rare in the academic written registers" however "more common in the written non-academic registers than in any other university register" (Biber & Barbieri, 2007, p. 281-282).

As well as studies on LBs in different registers, some studies compared professional and novice writings such as Cortes (2004) who conducted a study on the differences between professional and novice writers regarding bundles. The frequency and function of lexical bundles were compared in the written production of published authors and student essays in the history and biology. These expressions were frequently used in the corpora of published articles in those disciplines. However, the use of these expressions by the students in this study was extremely rare, even though they were all native speakers of English at various levels of their undergraduate years. Cortes' (2004) study presented that students could not transfer the lexical bundles into their written productions, although they were exposed to these expressions in the reading materials.

As a reason for this difference, Cortes suggested students may not have been formally instructed on both frequency and function of these expressions within the disciplines.

Similarly, Hyland (2008a) investigated the structures and functions of LBs (four-word lexical bundles) in the research articles, master theses, and doctoral dissertations from Applied Linguistics, Electrical Engineering, Business Studies and Biology disciplines. He aimed to reveal whether there was a disciplinary diversity in the frequencies and uses of LBs. He attempted to find out "the extent to which phraseology contributes to academic writing by identifying the most frequent 4-word bundles in the key genres of four disciplines" (Hyland, 2008a, p. 19). By the frequency of 20 times per million words, bundles were structurally and functionally classified. The study revealed a significant variation in the frequency of forms, structures, and functions of bundles across the disciplines by confirming the previous studies by Cortes (2004) and Biber (2006).

Hyland (2008b) found out significant differences between student and professional written performances in relation to the structure and functions of LBs identifying "the research articles, for example, contained far fewer clusters and far fewer different clusters overall…., they revealed more participants strings and included a far higher proportion of text-oriented clusters" (Hyland, 2008b, p. 59). In contrast, master theses revealed the opposite patterns. It was concluded that clusters play an important role in author experience and expertise at different levels.

In addition, Allen (2009) investigated the frequency and types of lexical bundles in the Active Learning of English for Science Students (ALESS) Corpus to find out the grammaticality, structural integrity, and functions of LBs in learner writing. The ALESS learner corpus comprise of 847 final research papers. Abstracts, introductions, methods, results, discussion, and conclusions sections have been included in the papers. The study revealed a striking convergence of LBs used between learners and professional writers. Besides, as students revised and edited one another's papers in peer conferencing and individual reviews, the grammatical accuracy was quite high. As for the grammatical structures of LBs used, noun phrase constructions which are -*NP + of*- were overused by the learners.

Some other studies focused on native and non-native academic writing. For example, Chen & Baker (2010) compared LBs used in native and non-native academic writing to find out problematic aspects in SLA regarding bundles. Three corpora were analyzed by using qualitative and quantitative data. It was aimed to reveal the similarities

and differences in using lexical bundles at different writing proficiency levels. Learner corpus contained writings from L1 Chinese learners of L2 English. The other two corpora had L1 writings from native academics and university students. The corpora were Freiburg-Lancaster-Oslo/Bergen (FLOB) corpus and British Academic Written English (BAWE) corpus. Significant similarities and differences between native and non-native academic writing were revealed. Using lexical bundles in non-native students' academic writing had more VP-based bundles and discourse markers than native academic writing, "which appears to be a sign of immature writing" (Chen & Baker, 2010, p.44) while native academic writing included more NP-based bundles and referential bundles. Chen & Baker (2010, p. 44) expressed that native student writing had the "control of cautious language in native professional writing", while non-native writers underused some high-frequent bundles of native academic writing and overused certain lexical bundles which native writers hardly used.

Another study on non-native academic writing by Wei & Lei (2011) examined the use of LBs in advanced Chinese EFL learners' academic writings. The corpus used included 20 doctoral dissertations from Applied Linguistics in the years between 2004-2009 and 120 published articles of six SSCI journals of Applied Linguistics in the years between 2004-2008. Four-word bundles have been investigated by 20 times per million words cut-off point frequency. According to the findings, advanced learner writers used much more lexical bundles and much more varied LBs in their writing than professional writers. When looked at the structural properties of these bundles, advanced learners used "similar number of prepositional phrases, noun phrases, be+ noun/ adjectival phrases and other structures of bundles to that of professional writers" (p. 164). As for the functional categories, advanced Chinese EFL learners produced a similar amount of research-oriented and text-oriented bundles with published writers. They used less participant-oriented bundles compared to professional writers. The researcher expressed it is not possible to learn all kinds of bundles in the natural language learning process. That is why, as Cortes (2004) points out, unconscious learning is not sufficient for learning these expressions. Confirming this view, Wei & Lei (2011) also suggested making learners 'notice' these recurrent word combinations.

Within the same line of reasoning, the use of LBs in advanced learner writing by L1 speakers of Swedish and undergraduate native speakers of linguistics have been analyzed by Adel & Erman (2012). They used both quantitative and qualitative data analysis. The corpus used was Stockholm University Student English Corpus (SUSEC)

of 325 essays which had one million words. Swedish and British undergraduate linguistic students at different proficiency levels were asked to contribute to the academic writing corpus. The study concluded that NNS use limited and less varied lexical bundles than NS. More varied lexical bundles used by NS were "in unattended 'this' constructions, existential 'there' constructions, hedges and passive constructions" (Adel & Erman, 2012, p. 90).

There were other studies conducted to create informative lists of LBs for pedagogical aspect of these expressions besides the studies on LBs in different registers and student-professional writing and native-nonnative academic writing. The first study belongs to Byrd & Coxhead (2010) investigating functions of LBs across different disciplines including Arts, Law, Commerce and Science. Each of the disciplines had seven subject areas. Researchers used the very same corpus compiled to develop the Academic Word List by Coxhead (2000). The corpus included 414 texts which were classified into three categories as short texts (2000-5000 running words); medium length texts (5000-10,000 running words) and long texts (over 10,000 running words). Each of seven sub-fields included 875,000 running words. Journal articles, course workbooks, book chapters, course notes and laboratory manuals constitute the corpus. Byrd and Coxhead (2010) compared and contrasted the frequency of each LB across disciplines by 20 times per million words cut-off point. The results indicated that all four disciplines have different bundles. That is why the list was reduced by choosing the bundles that consist at least 10% of each discipline. Only 35 bundles were on the remaining list. The researchers compared the 35 "shared and highly frequent bundles" (p. 39) with the lists of LBs published by Biber et al. (2004) and Hyland (2008a). Finally, a list of 21 lexical bundles was compiled by Byrd & Coxhead (2010) "that can be viewed by teachers and material writers as highly important and fairly stable across a variety of types of academic prose" (p.39).

Similarly, Simpson-Vlach & Ellis (2010, p. 478) also conducted a similar study by creating "an empirically derived and pedagogically useful list of formulaic sequences for academic speech and writing" called Academic Formulas List (AFL). The AFL has common lexical bundles in both academic spoken and written corpora after the examination of 2.1 million words of written and 2.1 million words of spoken academic discourse across various disciplines. The corpus included academic speech and writing corpora. While Michigan Corpus of Academic Spoken English (MICASE) and British National Corpus (BNC) files of academic speech constituted the spoken corpus, Hyland's

(2004) research article corpus and BNC files from academic disciplines made up the written corpus. The spoken corpus included texts from Biological Sciences, Physical Sciences, Social Sciences, Arts, and non-departmental/other. Written one included Social Sciences, Natural Sciences, Humanities and Arts, Technology and Engineering. The frequency cut-off was determined to be 10 times per million words. After that, 108 academic formulas were extracted depending on the n-gram length (3, 4, 5), frequency band (high, medium, and low) and mutual information (MI) band (high, medium and low). MI was used for collocation strength to address low psychological salience and pedagogical inutility of the lists of units in the lexical bundle research. What distinguishes this study is that lists compiled were presented to the twenty experienced English for Academic Purposes (EAP) instructors to rate the formulas in terms of formulaicity, having "a cohesive meaning or function as a phrase" and whether these formulas were worth teaching pedagogically (Simpson-Vlach & Ellis, 2010, p. 496). The study yielded a list of formulaic sequences that "are common in academic spoken and written language, and those that are special to academic written language alone and academic spoken language alone" (p. 487). Later, the formulas were grouped into according to their functions of referential, stance expressions and discourse markers. Also, this research contrasted with Hyland's (2008) research as it concluded that lexical bundles were not common to multiple disciplines, suggesting a strictly discipline-bound pedagogical approach to lexical bundles while giving importance to disciplinary variation. However, Hyland (2008) only examined four-word lexical bundles, while Simpson-Vlach & Ellis (2010) dealt with three-word lexical bundles. Hyland (2008) set a higher cut-off point of 20 times per million words, whereas Simpson-Vlach & Ellis (2010) used lower frequency cut-off point of 10 times per million words. That is why this research only included core lexical bundles common to all academic disciplines.

Martinez & Schmitt (2012) carried out a study to create a Phrasal Expressions List (Phrase List). In a pedagogical aspect, the need and place for formulaic sequences in the lexicon is significant. So, the researchers attempted to generate lists of recurrent word combinations which serve pedagogical purposes similar to the Academic Word List (Coxhead, 2000). They created a list of 505 lexical bundles at the end of the research.

As for the corpus-based studies on the use of lexical bundles of Turkish writers, studies were a few as can be seen from the Table 6.

Table 6.

Corpus-based Studies on Turkish Writers' Lexical Bundle Use

Researcher(s)	**Year**	**Focus of the Study**	**Corpus**
Karabacak & Qin	2012	Lexical bundles used by Turkish, Chinese, and American Students	University students' argumentative papers
Öztürk	2014	Lexical bundle use by Turkish and native English writers	MA/PHD theses, native writers' research articles
Muşlu	2015	Stance bundles in academic writing English, Turkish and Japanese learners	TICLE, LOCNESS JPICLE (Japanese International Corpus of Learner English)
Uçar	2017	Lexical bundles in the academic writing by Turkish and native English writers	15 scientific articles of native and Turkish advanced writers
Bal Gezegin	2019	Lexical bundles in Turkish writers' research articles	200 research articles

When the studies in the table presented examined, Öztürk (2014) examined bundles in Turkish and native English postgraduate students' writings in an academic discipline. The corpora used were made up of 150 texts from Turkish and English postgraduate students' both MA and PhD theses and native writers' published research articles between the years of 2003 and 2013. The cut-off point was set to 25 times per million words. The study categorized the bundles according to their structures, functions and frequency compared to the control corpus. The study revealed that Turkish postgraduate students used more frequent bundles than native students and writers. However, most of the lexical bundles were overused by Turkish postgraduate students.

Unlike these two studies Bal Gezegin (2019) and Öztürk (2014), Karabacak & Qin (2012) compared and contrasted the use of lexical bundles in the argumentative essays of Turkish, Chinese, and American writers. The argumentative papers belonged to Turkish sophomores, Chinese sophomores, and American first-year students as a learner corpus. Besides, New York Times articles were selected as a reference corpus. The study concluded that despite simple exposure, advanced English learners had difficulty in acquiring some lexical bundles. As a solution to this issue, the researchers suggested explicit teaching of lexical bundles to improve their acquisition procedure.

Muşlu (2015) investigated the stance lexical bundles in the argumentative essays by native English speakers and Turkish and Japanese EFL learners. The corpora used

included LOCNESS, TICLE (Turkish International Corpus of Learner English) and JPICLE (Japanese International Corpus of Learner English). The lower cut-off points of 3 per million words have been set because of the amount of data. The researcher compared the stance bundles in these three corpora and categorized them according to their functions and structures, which vary in each group. It was concluded that native speakers use less but various lexical bundles. While lexical bundles have been used higher in Japanese and Turkish corpora, lexical variety is lower than the native data.

Likewise, another study examined the most frequent lexical bundles in published academic articles of Turkish and native speakers of English (Uçar, 2017). The data comprised two corpora which are 15 scientific articles of NS and 15 scientific articles of Turkish advanced writers. Three-word lexical bundles were extracted using frequency criteria by the frequency cut-off point of 40 times per million words. Then, the bundles were categorized in terms of their structures and functions. The study concluded that Turkish writers underused some bundles and used less various bundles in their academic productions compared to the native writers.

Another study examined the use of lexical bundles in the research articles written by Turkish writers was examined by Bal Gezegin (2019) in the first study. A corpus comprising 200 published research articles in six different disciplines including Economics, Medicine, Psychology, Education, History and Sociology was compiled. In the study, four-word lexical bundles were analyzed with a frequency cut-off point of 20 times per million words. Bal Gezegin (2019) identified ninety-nine lexical bundles in Turkish Scholars Research Articles Corpus (TSRAC). It was revealed that '*on the other hand, the end of the, as well as the, in the case of* and *one of the most*' are the most frequent lexical bundles in TSRAC. Also, the bundles were classified both structurally and functionally.

To conclude, the review of literature in the field displayed many studies on the use of lexical bundles. The studies have demonstrated that non-native academic writers have difficulty in acquiring and producing lexical bundles in a native-like manner. Similarly, studies revealed a significant difference between native and non-native writing. To exemplify, NNS overused or underused some lexical bundles in their academic writing and used more limited and less varied lexical bundles. Besides, the bundles produced by students differ from the bundles used by professional writers (Cortes, 2004; Hyland, 2008b; Allen, 2009; Chen & Baker, 2010; Adel & Erman, 2012; Öztürk, 2014).

2.7.4. Major Studies on Pedagogical Aspect of Lexical Bundles

The lexical bundles are not acquired naturally as mentioned above (p. 28-29) and simple exposure to the lexical bundles is not sufficient for learners to produce the bundles actively (Cortes, 2004, 2006; Wei & Lei, 2011; Karabacak & Qin, 2012). It is also an issue for advanced non-native English learners and second language learners to acquire these expressions (Bishop, 2004; Karabacak & Qin, 2012). That is why formal instruction of lexical bundles may be one solution to make non-native writers notice the lexical bundles in their writing.

There are several other studies which focused on enhancing and noticing the acquisition of formulaic sequences. One of the leading studies belongs to Li & Schmitt (2009) who carried out a longitudinal study on the formulaic sequences in L2 academic writing. The participant of the study had been studying English for 10 years and the researchers analyzed 8 writing assignments along with a dissertation for lexical bundle usage. The participant was also interviewed after each assignment. The study revealed that the participant learned 166 new bundles and gained confidence relating to the use of bundles thanks to the treatment. However, researchers expressed that the participant used a restricted number of bundles which made her writing to be non-nativelike. Another significant finding of this study was that the formulaic sequences were acquired incrementally. This implies a resemblance to the acquisition of single words (Nation, 2001).

Likewise, the effects of three types of explicit instruction on formulaic sequences were examined by Nguyen (2014). The participants were Mandarin-speaking graduate students with a TOEFL score ranging between 90 and 110. The participants were divided to four groups with one control and three experimental groups. The treatment groups were provided with three types of intervention; Input Enhancement together with Explicit Instruction; Collaborative Gap-filling Tasks and Spot the Different tasks. The control group received no instruction on formulaic sequences. The data collection instruments were Vocabulary Knowledge Scale, an awareness test as pre-tests, a Cued gap-filling Test, a multiple-choice test, and the same awareness test used as immediate and delayed post-tests. The study revealed that three treatment groups displayed performance significantly higher than the control group. As a result, experimental groups had a more positive effect on the learners' receptive and productive skills of acquiring formulaic sequences than the control group. It was also concluded that the most useful one was the gap-filling test among three types of treatment for the acquisition of productive

knowledge of formulaic sequences. While direct instruction on formulaic sequences made learners retain meaning most effectively, explicit strategy teaching helped learners notice formulaic sequences.

In their experimental research, Jones & Haywood (2004) investigated the effectiveness of explicit instruction of formulaic sequences in terms of awareness, accuracy, and production of the sequences. The researchers also wanted to improve learners' strategies in an EAP context. The participants were 21 students from the Centre for English Language Education at the University of Nottingham, and the study took place for ten weeks. The control group consisted of 11 participants, whereas the treatment group had 10 participants. The researchers specified the target formulaic sequences in academic writing from Biber et al.'s (1999) study, considering the usefulness and specific language functions. In the experimental group, the specified expressions were provided to learners with reading activities including highlighted sequences in reading texts, concordance lines and corpus extracts along with writing activities. At the same time, interviews and classroom observations were carried out to reveal students' reactions. In order to reveal learning performances, a pretest and a posttest were given. As a result, the study concluded that most of the students in the treatment group performed a significant increase in the awareness of the formulaic sequences. While a slight improvement in the learners'-controlled production of formulaic sequences measured by C-test was observed, there was not any improvement in learners' free production of formulaic sequences. The findings of the study were not in line with the aforementioned research to produce formulaic sequences (Schmitt et al., 2004; Nguyen, 2014). The researchers explained this with some limitations of the study stemming from time and curriculum constraints (only two weeks were available between two writing tests) and the fact that repeated exposure treatment included only noticing and retrieval activities.

Some other studies measured the effect of explicit instruction on receptive and productive knowledge of bundles in ESL and EFL contexts. For instance, Kazemi, Katiraei & Rasekh (2014) investigated whether formal instruction would develop Iranian advanced Teaching English as a Foreign Language (TEFL) students' writing skills and their attitudes towards the teaching of lexical bundles. The participants were students of Master of Arts (MA) in Applied Linguistics. The researchers used pre-test, post-test, and a questionnaire in order to reveal students' attitudes towards the treatment. The treatment period lasted for a month. The participants were asked to write passages before and after the treatment period. In the treatment, the frequency, use and functions of bundles that

are specific to the field of applied linguistics were taught to the participants. The findings revealed that students improved their writing skills despite a very short duration of the treatment. The study also revealed that even though the participants were the advanced level of students, they were not familiar with the bundles and did not have any training related to the use of lexical bundles. The study underlined the importance of integrating these expressions into foreign language teaching process in conclusion.

In a similar vein, to investigate the implementation of the explicit instruction of lexical bundles to improve the writing skills of EFL learners, Latifi & Afraz (2015) carried out a study with 50 Iranian pre-intermediate students. The participants were divided into experimental and control groups. As the data collection instruments, a writing test as a pre-test and another writing achievement test as a post-test were administered. While the experimental group was given an explicit instruction of target lexical bundles, the control group were given a placebo instruction to improve their writing skills. The study concluded that explicit instruction helped students improve their writing skills.

Another study focused on teaching lexical bundles to university students in a writing-intensive history class (Cortes, 2006). She taught five different 20-minute micro-lessons for ten weeks. The participants were English native speakers. Among the exercises, there were contextualized examples from articles and filling the blanks activities. The reactions and thoughts of the students on the use of lexical bundles were discussed during the lessons. According to the findings of this study, no differences between pre-post instruction were revealed on the production of lexical bundles. However, an awareness was developed on these multi-word combinations. As for the researcher, the reason may be the length of micro-lessons since they were not sufficient for students to augment the productive knowledge of lexical bundles, or the activity types, which may not have been appropriate for learners to improve their production of these expressions in their writing.

Similar to the studies of Jones & Haywood (2004) and Cortes (2006), Čolović-Marković (2012) also focused on the explicit instruction of formulaic sequences on second language writing. The aim of the study was to reveal the effect of explicit teaching of sequences in controlled (c-tests) and in uncontrolled situations (essays). Also, it was aimed to find out students' strategies used while producing sequences in ESL writing. A quasi-experimental design had been used with experimental and control groups. The participants were in the writing classroom of the university intensive English program.

While the experimental group received explicit instruction of formulaic sequences through classroom discussion and reading to raise their awareness, the control group did not receive any kinds of explicit instruction, however, they focused on writing-oriented activities. Both quantitative and qualitative data were collected through the 8-week. The study revealed that the explicit instruction of formulaic sequences significantly affected students' performances in terms of producing academic formulaic sequences in a controlled situation and topic-induced formulaic sequences both in controlled and uncontrolled situations. No effect was observed on the production of academic formulaic sequences in an uncontrolled situation. The findings of the study were consistent with the findings of earlier research, including Jones & Haywood (2004), and Cortes (2006). The researcher also noted that she tried to eliminate the limitations factors in studies of Jones & Haywood, (2004) and Cortes (2006). The students were exposed to more practice and were more motivated.

When the importance of a pedagogical aspect of lexical bundles is considered as mentioned above, several teaching activities and lessons for the acquisition of these bundles have been designed by Neely & Cortes (2009) who examined the frequency of 5 topic-introducing and discourse organizing bundles revealed by Biber et al. (2004) and Nesi & Basturkmen (2006) in the academic lectures of instructors and students in the Corpus of Academic Spoken English (MICASE) and the teaching applications of these five bundles (*if you look at, a little bit about, a little bit of, I want you to, and I would like you*) in academic lectures. For each of the five topic-introducing lexical bundles were searched to reveal the frequency and the functions of these bundles. Later, there were compared with the research of Biber et al. (2004) and Nesi & Basturkmen (2006). It was revealed that lexical bundles could be used in a variety of academic lectures "while a lexical bundle can have a primary function, the same bundle can be used for different functions across the span of a lecture" (Neely & Cortes, 2009, p. 29). The researchers designed corpus-based lessons and materials for ESL/ EFL classrooms using corpora and concordance programs since these activities can be "effective teaching and learning tools when proper planning and instruction takes place" (Neely & Cortes, 2009, p.30). The aims of the lessons were to raise students' awareness, explicitly introduce lexical bundles and their functions, make students become familiar with the form and functions of lexical bundles. It was concluded that instructors ought to teach bundles by presenting their functions in context.

Lastly, Ergin (2013) investigated the effects of explicit instruction of formulaic

sequences on Turkish EFL learners' use of formulaic sequences in their overall writing performance. The participants of the study were English Language Literature students with upper-intermediate levels. Four-week treatments were included in the study. The instruments used were pre and post-tests. Also, content analysis was conducted by separating accurate and inaccurate discourse markers used. As for the results, when post-tests of the students were compared to pre-tests, there has been a significant increase in the number of formulaic sequences that is used. Besides, the formal instruction of formulaic sequences developed learners' usage of formulaic language in their written productions.

As can be seen above, the effects of explicit instruction on teaching lexical bundles/formulaic sequences on learners' receptive and productive writing abilities have been investigated in the studies. While some studies revealed a statistically significant development in the learners' knowledge of bundles both receptively and productively (Schmitt et al., 2004; Kazemi, Katiraei & Rasekh, 2014; Nguyen, 2014), other studies indicated learners' greater awareness of lexical bundles with no significant improvement on the production of those bundles in writing skills in uncontrolled situations (Jones & Haywood, 2004; Cortes, 2006; Čolović-Marković, 2012). Some limitations including short duration of micro-lessons and not appropriate activities for the learners (Cortes, 2006); time, curriculum (only two weeks were allocated between two writing tests) and repeated measure constraints (Jones & Haywood, 2004), and the necessity of more exposure along with more practice (Čolović-Marković, 2012) have been observed in these studies.

To sum up, many studies relating the LBs in different language backgrounds and corpus linguistics in the existing literature have been presented in this section of the dissertation.

CHAPTER III

METHODOLOGY

3.1. Introduction

Two general approaches exist in corpus studies: ‘corpus-based’ and ‘corpus-driven’ (Tognini-Bonelli, 2001). ‘Corpus-based’ research is a method that involves studying language through large collections of real-life language use, known as corpora. This approach allows researchers to analyze patterns, frequency, and variations of linguistic features in a systematic and empirical way. As the main purpose is to analyze the systematic patterns of use, corpus-based studies of formulaic language require the researcher to pre-select expressions. After that, the corpus is analyzed in order to discover how these expressions are used, for instance, by learners (Moon, 1997).

Adapting an inductive approach, ‘corpus-driven’ research assumes that the linguistic constructs emerge by analyzing a corpus. The corpus-driven approach is different from the standard practice of linguistics since it makes minimal assumptions about the linguistic constructs by analysis. Traditional linguistic analysis often relies on the intuition of the linguist, especially for syntax. In contrast, corpus-driven approaches are based on empirical evidence gathered from large collections of actual language use. As we implied before, traditional linguistic methods often start with a hypothesis or theory that is then tested or explored. This is a deductive approach. Corpus-driven approaches, on the other hand, are inductive. They start with the data and look for patterns, which can then lead to the formation of hypotheses or theories. Stemming from the nature of the data analysis it offers, corpus-driven approaches place a lot of emphasis on frequency and patterns of language use. This is different from traditional linguistics, which may focus more on the structure and rules of language. Moreover, it allows for the study of language variation and change over time or across different groups of speakers. This is less common in traditional linguistics, which often aims to describe an idealized, homogeneous version of a language. It can also provide insights into pragmatics and discourse, areas that are often less accessible to traditional methods due to the need for larger contexts of language use. While traditional linguistics often involves qualitative analysis, corpus-driven approaches allow for both quantitative and qualitative analysis of language.

In brief, corpus-driven analysis assumes only the existence of words and co-occurrence patterns among words by analyzing the corpus. And this is the basis for subsequent linguistic descriptions. In a corpus-driven analysis, the "descriptions aim to be comprehensive with respect to corpus evidence" (Tognini-Bonelli 2001: 84), so that even the 'linguistic categories' are derived "systematically from the recurrent patterns and the frequency distributions that emerge from language in context" (Tognini-Bonelli 2001: 87). Previous studies of 'lexical bundles' in spoken and written registers have adopted this approach (Biber et al. 1999: Biber et al., 2004).

3.2. Frequency Approach

The research on formulaic language generally falls into two categories in terms of methodology. The first one is Nesselhauf's (2004) phraseological approach implementing top-down methods to classify formulaic language by using a specific criterion like semantic or syntactic composition. The second approach is bottom-up corpus-driven approach (Sinclair, 1987) which is an inductive approach. It can also be referred to as frequency-based (Nesselhauf, 2004) approach. In this type of research, corpus analysis and frequency cut offs are used to identify formulaic sequences and these expressions are classified according to their discourse functions. Our study aims to use the corpus-driven approach (Sinclair, 1987) which is also referred to as frequency-based (Nesselhauf, 2004) approach.

Using frequency as the basis is useful in the way that patterns of use are identified which "otherwise often go unnoticed by the researchers" (Biber et al., 2004, p. 376). This approach is also pivotal for the MWIs because they are the reflection that a sequence of words is stored and used as a prefabricated chunk. Higher frequency sequences tend to be stored as unanalyzed chunks than lower frequency sequences (Biber et al., 2004). In other words, sequences of words that occur more frequently within a particular language tend to be stored and processed as integral units, known as "chunks", without undergoing additional dissecting or scrutinizing procedures. Conversely, sequences of words with lower frequencies do not receive the same treatment and are subject to further analysis. Nearly every theoretical description of the corpus-driven approach includes mention of frequency, as in the "linguistic categories" are derived "systematically from the recurrent patterns and the frequency distributions that emerge from language in context" (Tognini-Bonelli, 2001, p. 87) in a grammar pattern, "a combination of words occurs relatively frequently" (Hunston & Francis, 2000, p. 37).

3.3. Research Design

The current study investigates the use of 3- and 4- lexical bundles in the Turkish International Corpus of Learner English (TICLE) learner corpus containing the argumentative essays by Turkish EFL learners. According to a study by Biber et al. (1999), it was found that 3-word bundles, which refer to sequences of three consecutive words, occur at a frequency that is ten times higher compared to 4-word bundles, which are sequences of four consecutive words. Furthermore, the study indicated that, on average, these 3-word bundles were found to occur approximately 25 times per million words (pmw).

With a few exceptions, majority of the studies on lexical bundles focused on 4-word LBs and even longer bundles. However, Cortes (2021) recommends exploring frequent 3-word expressions for pedagogical purposes since they are more frequent, and they are not all part of the 4-word bundles. This study particularly examines the most frequent lexical bundles, their structures, and functions in the written performances of Turkish EFL learners. Our aim is to provide an understanding of uses and differences of lexical bundles. This is pivotal as these expressions are "defining markers of fluent writing and are important for the development of writing" (Coxhead & Byrd, 2007, p. 135). The steps for the analysis administered in conducting the study are presented in Table 7 below.

Table 7.

Research Types and Stages of the Study

Stages	**Process**	**Research Type**
Stage 1	Automatic generation of frequency lists	Descriptive analysis
Stage 2	Selection of meaningful LBs manually	Corpus-based approach
Stage 3	Application of statistical analyzes across corpora	Quantitative corpus analysis
Stage 4	Analysis of LBs structurally and functionally	Descriptive analysis
Stage 5	Comparison of LBs between 2 corpora	Interpretative
Stage 6	Comparison of semantic domains	Descriptive analysis

The table summarizes the stages of the study along with the research types employed. Firstly, all the 3 and 4-word sequences have been extracted from LOCNESS and TICLE using Antconc (Anthony, 2020) software as an analysis tool. Among these

sequences, meaningful multiword items have been identified manually in both corpora. The overlapping bundles as well as repetitive and erroneous ones have been extracted from the list. When analyzing the data, LBs including abbreviations have been merged with the ones with no abbreviation so that these items would not stand as different word groups. To exemplify, *I do not want* and *I don't want*, and *it is not* and *it's not* have been merged and counted as a single bundle. The frequencies of LBs and structural and functional classifications are summed up in the analysis section. Secondly, a cut-off point has been set after the selection process (Cortes, 2004) as explained above on page 57.

All identified bundles in each corpus have been categorized based on their structural and functional characteristics. The characteristics of these most frequent bundles in both corpora have been sorted adapting Biber et al.'s (1999; 2003; 2004) structural and functional classification of lexical bundles framework. Additionally, similarities and differences of the frequent 3- to 4-word formulaic productions between Turkish EFL learners and native English speakers have been revealed. The native speaker corpus used in the present study is Louvain Corpus of Native English Essays (LOCNESS) which is a native reference corpus containing American Students' written essays (Granger et al., 2009). After revealing the interpretative contrastive analysis, log-likelihood analysis has been carried out in order to obtain the statistical significances based on the normalized sizes of the corpora. In the final step, semantic domains of these expressions in both corpora have been determined. UCREL Semantic Analysis System (USAS) and its semantic tag sets have been used used via Wmatrix (Version 4) semantic analysis platform (Rayson, 2008) to identify the most salient features of the multiword preferences in both corpora.

3.4. Data Collection Tools and Participants

In this study, Turkish International Corpus of Learner English (TICLE), a sub-corpus of the International Corpus of Learner English (ICLE), and the Louvain Corpus of Native English Essay (LOCNESS) have been used to extract the data. Each corpus consists of argumentative essays. Aijmer (2009) asserts that these types of essays allow for both discourse-oriented (cohesion, coherence, argumentative patterns, etc.) exploration and lexical and grammatical exploration. The breakdowns and components of these corpora used is given in the following sections.

3.4.1. The Turkish International Corpus of Learner English (TICLE)

TICLE (Turkish International Corpus of Learner English) has been selected as the main corpus. TICLE is learner corpus compiled from higher intermediate to advanced learners of English as a foreign language (B2 according to CEFR). The corpus architecture was designed by the Centre for English Corpus Linguistics at the University of Louvain. In addition, numerous partner universities around the world also collaborated in the collection process. The first version appearing in 2002, the ICLE corpus contained 2.5 million words by learners from 11 mother tongue backgrounds. Later, the second version with 3.7 million words representing 16 mother tongue backgrounds including Turkish, German, Dutch, Finnish, and Spanish were released in 2009 (Granger et al., 2009). The third version of the ICLE, around 5 million words, comprises English data by learners from the following L1s: Brazilian Portuguese, Hungarian, Greek, Persian (Iran), Lithuanian, Macedonian, Korean, Pakistani and Serbian.

Considering the highly heterogeneous nature of learner language, strict criteria were set at the onset of the Project ICLE. The corpus encompassed various variables that have the potential to influence learner performances. These variables are task variables (medium, genre, field, length, topic, and task setting) and learner variables (age, gender, mother tongue, region, other foreign languages, staying in English-speaking country, learning context, and proficiency level). Participants have been offered a wide range of topics for their argumentative essays. Top ten essay topics chosen by the learners are presented in Table 8.

Table 8.

Top Ten Essay Topics in ICLE

Essay topics	**Number of essays**	**Country of origin**
Some people say that in our modern world, dominated by science, technology and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?	491	29% Bulgarian
Most university degrees are theoretical and do not prepare students for the real world. They are therefore of very little value.	249	22% Turkish
Poverty is the cause of the HIV/AIDS epidemic in Africa.	243	100% Tswana
Marx once said that religion was the opium of the masses. If he was alive at the end of 20th century, he would replace religion with television.	237	19% Russian
The prison is outdated. No civilized country should punish its criminals; it should rehabilitate them.	176	32% Tswana
Discuss the advantages and disadvantages of banning smoking in restaurants.	156	100% Chinese
Discuss the advantages and disadvantages of using credit cards.	149	100% Chinese
Feminists have done more harm to the cause of women than good.	139	100% Russian
In the words of the old son “Money is the root of the evil”.	133	22% Russian
In his novel “Animal Farm”, George Orwell wrote “All men are equal: but some are more than others”. How true is this today?	127	39% Bulgarian

Source: Granger et al., 2009, p. 6-7

As can be seen from the table, ICLE is comprised of argumentative essays with a wide range of topics from feminism to HIV/AIDS epidemic in Africa. The corpus consists of academic writing that predominantly adopts an argumentative approach, with each sub corpus containing approximately 200,000 words (Kilimci and Can, 2008; Can, 2009).

Each sub-corpus has been divided into many batches with variable numbers of essays in ICLE. A certain coding system is applied to identify them. Each batch of the essays is given a 5-character code. The first two codes are for nationality (e.g., TR for Turkish). These two letters are followed by the institution code (e.g., CU for Çukurova University). However, if more than one batch exists within an institution, a single letter

code is given. For example, TRCU1 refers to Turkish Sub-corpus, Çukurova University, 1st batch. The following table summarizes the breakdown of TICLE corpus. Text types of the data are Argumentative (A).

Table 9.

The Breakdown of TICLE Corpus

Batch	**Essay codes**	**Text types (Argumentative)**	**Number of essays**	**Number of words**
TRCU1	TRCU1001-1177	177	177	128,297
TRKE2	TRKE2001-2072	72	72	50,228
TRME3	TRME3001-3031	31	31	21,007
TOTAL		280	280	199,532

Source: Can, 2009, p. 26.

As seen from the table, TICLE contains 280 essays with 199,532 words. Following the design criteria, TICLE, as a sub corpus, also consists of argumentative essays. The learners are young adults (university undergraduates), they are in B2 proficiency level and are still learning English as a Foreign Language (EFL) rather than as a Second Language (ESL).

The present study drew upon corpus data in TICLE collected from three participating academic institutions, namely the Cukurova University, the University of Mustafa Kemal and Mersin University gathered the essays (Can, 2009). The collected essays exhibited an average word count of 713, reflecting the length of the compositions produced by the participants. Additionally, the average age of the participants was recorded as 22,08 and 81% of the participants were female (Kilimci and Can, 2008).

3.4.2. Louvain Corpus of Native English Essays (LOCNESS)

The native speaker reference corpus known as LOCNESS was compiled by the Centre for English Corpus Linguistics at the Catholic University of Louvain, Belgium under the authorship of Granger (1998). It became publicly accessible in 1998, offering a valuable resource for linguistic research. The essays contained within the corpus were amassed during the period spanning from 1991 to 1995. Comprising a comprehensive collection of British students' A-level essays (60,209 words), British university students' essay (95,695 words), and American university students' essays (168,400 words), the

LOCNESS corpus encompasses a diverse range of academic compositions. Of the American students' essays, 149,574 words are dedicated to argumentative essays while 18,826 words represent a literary-mixed genre. Consequently, the overall word count of LOCNESS corpus amounts to 324,304 words. In terms of the essay length, it is noteworthy that the average falls around 500 words, aligning closely with the ICLE corpus in this regard.

Regarding the text types encompassed within the corpus, it comprised examination papers, timed essays, and free essays. It is important to note that while certain timed and free essays incorporated the use of reference tools, examination papers did not permit their usage.

Within this dissertation, the selected corpus comprises American university students' essays (149,574 words), which were chosen due to the similarity in age distribution with TICLE corpus participants, as well as their shared characteristic of being exclusively argumentative in nature. The age range of the participants falls between 17 and 28 years. The essay topics encompass a wide array of subjects, including but not limited to sex equality, water pollution, gender roles, teenagers, homosexuality, nuclear power, yoga, controversy in the classroom, prayer in schools, violence, and AIDS. To provide a comprehensive overview of the corpus utilized in this specific study, the following table succinctly summarizes its key characteristics.

Table 10.

LOCNESS Selected Components

Institutions	**Codes**	**Text types**	**Number of essays**	**Number of words**
Marquette University	ICLE-US-MRQ	Argumentative essay	46	54,285
Indiana University, Indianapolis	ICLE-US-IND	Argumentative essay	28	13,454
Presbyterian College, South Carolina	ICLE-US-PRB	Argumentative essay	6	12,447
University of South Carolina	ICLE-US-SCU	Argumentative essay	53	52,885
University of Michigan	ICLE-USMICH	Argumentative essay	43	16,502
Total			**175**	**149,574**

While the overall word count for American students' essays amounts to 168,400, it is noteworthy that only the argumentative essays, which constitute 149,574 words, have been utilized for this study. The remaining 18,826 words belong to the literary-mixed essays and have been excluded from analysis. The total number of American students' essays included in the study is 175, with a cumulative word count of 149,574. The participating universities in this research endeavor encompass Marquette University, Indiana University, Marquetta University, University of Michigan, University of South Carolina, and Presbyterian College.

3.5. Data Analysis Tools and Statistics Used

Within this dissertation, a combination of descriptive and quantitative research methodologies has been employed to address the research questions posited earlier. To facilitate data analysis, various corpus analysis tools have been utilized. The integration of these corpus analysis tools provides a comprehensive framework for exploring and interpreting the data in this dissertation.

3.5.1. Cut-off Point

As mentioned in the previous sections, formulaic sequences are “recurrent expressions, regardless of their idiomaticity, and regardless of their structural status” as Biber et al. (1999, p. 990) define. These expressions have to meet a set of criteria to be considered as bundles or multiword items such as identifying these items by corpus analysis tools (e.g., AntConc, Lancsbox...etc.) by using the criterion of frequency cut-off; and assigning functions to the sequences identified by frequency and range criteria.

The frequency cut-off point varies in each study. The actual frequency cut-offs in lexical bundle research vary from ten occurrences per million words (Biber et al., 1999; Biber, 2006), to forty occurrences (Biber, Conrad, & Cortes, 2004). Lower cut-offs may be used in smaller corpora (Biber et al., 1999) and determining the frequency cut-offs base on the earlier researchers rather than a statistical and empirical set of standards. Some other researchers such as Hyland (2008) employed a percentage criterion that required a sequence to appear in at least 10 percent of the texts in a corpus. Additionally, these MWIs have to appear in at least three to five different texts (Biber & Barbieri, 2007). This way, focusing on idiosyncratic uses by the authors is prevented.

For the aforementioned reasons, the cut-off point for the TICLE corpus has been

determined as 10 however, the cut-off point for the LOCNESS corpus has been 7 in this dissertation. The cut-off points in the analysis of both corpora have been determined with the aim of increasing the accuracy and reliability. Also, the sequences which were seen in at least 3 different texts have been chosen during the identification of sequences in order to make the analysis more accurate. That is why, MWIs which have been used 20 times in three different texts have been included.

3.5.2. Type / Token Ratio

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

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

T/t is obtained by dividing the type count by the token count, which is always <=1. While a high T/t score indicates a high degree of lexical variation, a low T/t indicates a low degree of lexical variation. T/t can also be expressed as a percentage, multiplying the ratio by 100 (Thomas, 2005). T/t ratio of MWIs represents the percentage of each item within all words in a corpus, which is the actual number of MWIs that fall into per 100 words. In our study, we have utilized the very same formula in our data analysis.

Tokens are the number of words in a text. Yet, many of these tokens are repeated in corpora (Williamson, 2009). To exemplify, the tokens with their frequency occurrences can be seen below.

rank	word	freq	rank	word	freq	rank	word	freq	rank	word	freq
1	we	6	17	asleep	1	33	impressions	1	49	seem	1
2	and	5	18	at	1	34	instant	1	50	so	1
3	them	5	19	beliefs	1	35	is	1	51	the	1
4	are	3	20	brief	1	36	just	1	52	then	1
5	can	3	21	but	1	37	later	1	53	things	1
6	they	3	22	call	1	38	metaphor	1	54	thinking	1
7	to	3	23	coming	1	39	mull	1	55	this	1
8	again	2	24	concepts	1	40	notions	1	56	thoughts	1
9	as	2	25	described	1	41	occasions	1	57	times	1
10	in	2	26	endure	1	42	opinions	1	58	values	1
11	on	2	27	fall	1	43	other	1	59	variously	1
12	a	1	28	going	1	44	over	1	60	views	1
13	act	1	29	handle	1	45	perceptions	1	61	well	1
14	all	1	30	have	1	46	put	1	62	what	1
15	an	1	31	however	1	47	refer	1	TOTAL		87
16	aside	1	32	ideas	1	48	return	1			

Figure 3. A Sample figure for type-token ratio

Source: Williamson, 2009

In Figure 9, the token *again* occurs two times and the token *we* occurs six times. When the number of words is counted, a total of 87 words exists. While there are this total of the 87 tokens, 62 ‘types’ exist. The T/t score for this sample is calculated as:

(62/87) * 100 = 71.3%

According to this calculation, 71.3 of the words belong to a certain type of group in 100 words. If there are more types when compared to the number of tokens, more varied vocabulary exists. Therefore, lexical variety is greater.

3.5.3. Log-Likelihood

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

The observed and expected values are compared with LL in two datasets. While the observed values are the actual frequencies extracted from corpora, expected values are the frequencies that one would expect if no factor other than chance were affecting the values. The greater the difference between the observed and the expected values, the

less likely it is that the difference has arose by chance. Log-likelihood is calculated by constructing a contingency (frequency) table as follows:

	Corpus 1	Corpus 2	Total
Frequency of word	a	b	a+b
Frequency of other words	c-a	d-b	c+d-a-b
Total	c	d	c+d

Figure 4. Contingency figure for log-likelihood
Source: Rayson and Garside, 2000

The value 'c' corresponds to the number of words in corpus one, and ’d’ corresponds to the number of words in corpus two (N values). The values 'a' and 'b' are called the observed values (O), the expected values (E) are calculated according to the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

In the formula, N1 = c, and N2 = d. So, for this word, E1 = c*(a+b) / (c+d) and E2 = d*(a+b) / (c+d). The calculation for the expected values takes account of the size of the two corpora, so we do not need to normalize the figures before applying the formula. We can then calculate the log-likelihood value according to this formula:

$$-2\ln\lambda = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

This equates to calculating log-likelihood G2 as follows: G2 = 2*((a*ln (a/E1)) + (b*ln (b/E2))). Contingency table as presented in Figure 10 has two rows and two columns and has 1 degree of freedom. In LL, critical values with 1 df are 3.84, 6.64 and 10.83 for the significance levels of 0.05, 0.01 and 0.001. For example, 28.841 (1 df) is greater than 10.83 for the significance level 0.001. If the log-likelihood for the result is greater than 6.63, the probability of the result- i.e., the difference between the two corpora - happening by chance is less than 1%. Thus, we can be 99% certain that the result actually means

something. This is usually expressed as $p < 0.01$. If the log-likelihood is 3.84 or more, the probability of it happening by chance is less than 5% meaning 95% certainty of the result. This is expressed as $p < 0.05$.

When we obtain the frequencies for our corpus searches, we make the results comparable using frequencies by percent, or by per million words as mentioned above. This is called normalizing the frequencies. However, normalized scores do not mean that what we have is significant. Based on the information provided, we carried out our data analysis using the LL mentioned above.

3.5.4. The AntConc

In the field of corpus linguistics, researchers rely on computer software dedicated corpus tools to analyze the data, conduct comparisons between corpora, and derive empirical results. We have decided to use the AntConc (Anthony, 2020) which serves as a freeware corpus analysis toolkit designed specifically for concordancing and text analysis purposes. The AntConc is a toolkit that is used to create frequency lists, search concordances, identify collocations, find clusters (frequency patterns of word sequences) or n-grams (sequences of n words within the corpora or documents) (Anthony, 2020). The interface of the AntConc may be seen in Figure 3 below.

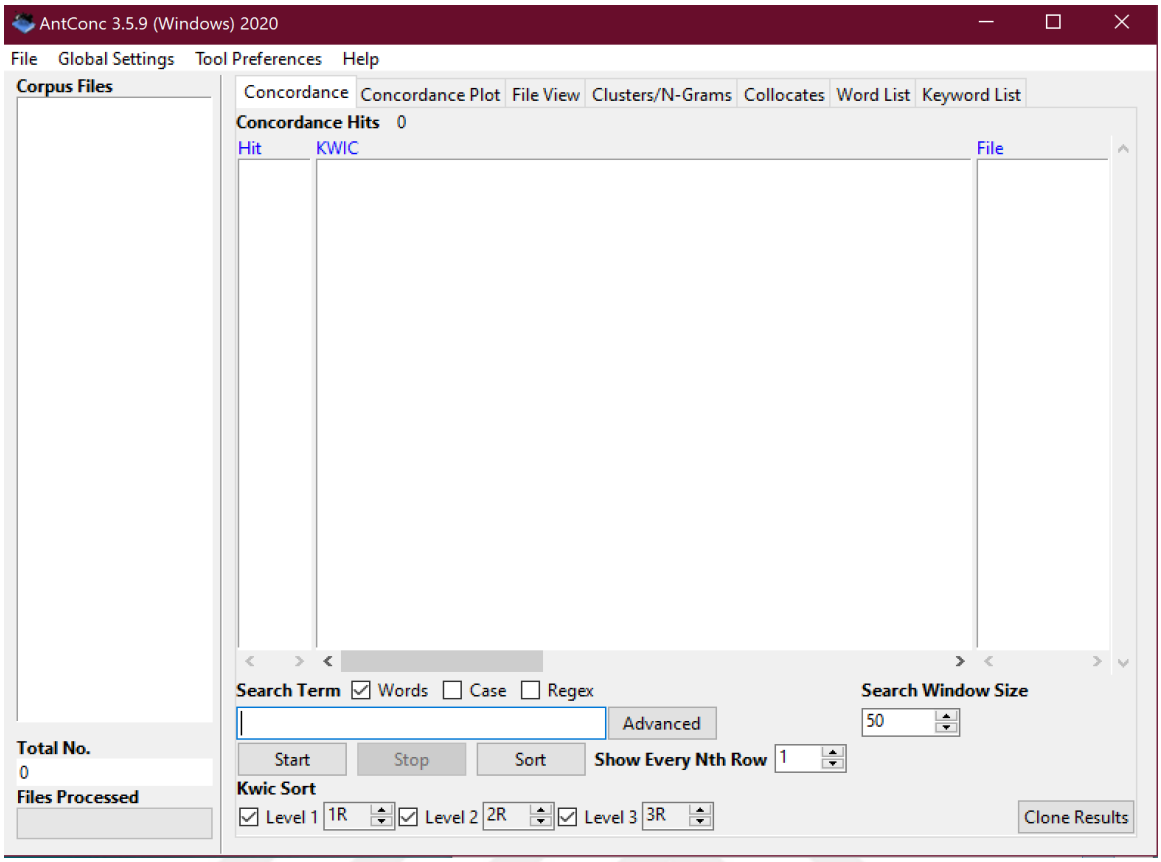


Figure 5. A Screenshot of AntConc interface

Source: Anthony, 2020

In the study, we have employed the Clusters/N-grams tool of AntConc for our analysis. This tool enables the exploration and search of clusters and n-grams using the provided search option. Researchers can customize searches based on the desired n-gram size, frequency, and range parameters. The discovered phrases can be arranged and presented in alphabetical, range, word, or frequency order, allowing for convenient analysis and interpretation. Figure 4 presents the Clusters/N-grams tool, providing a visual representation of its interface and functionality.

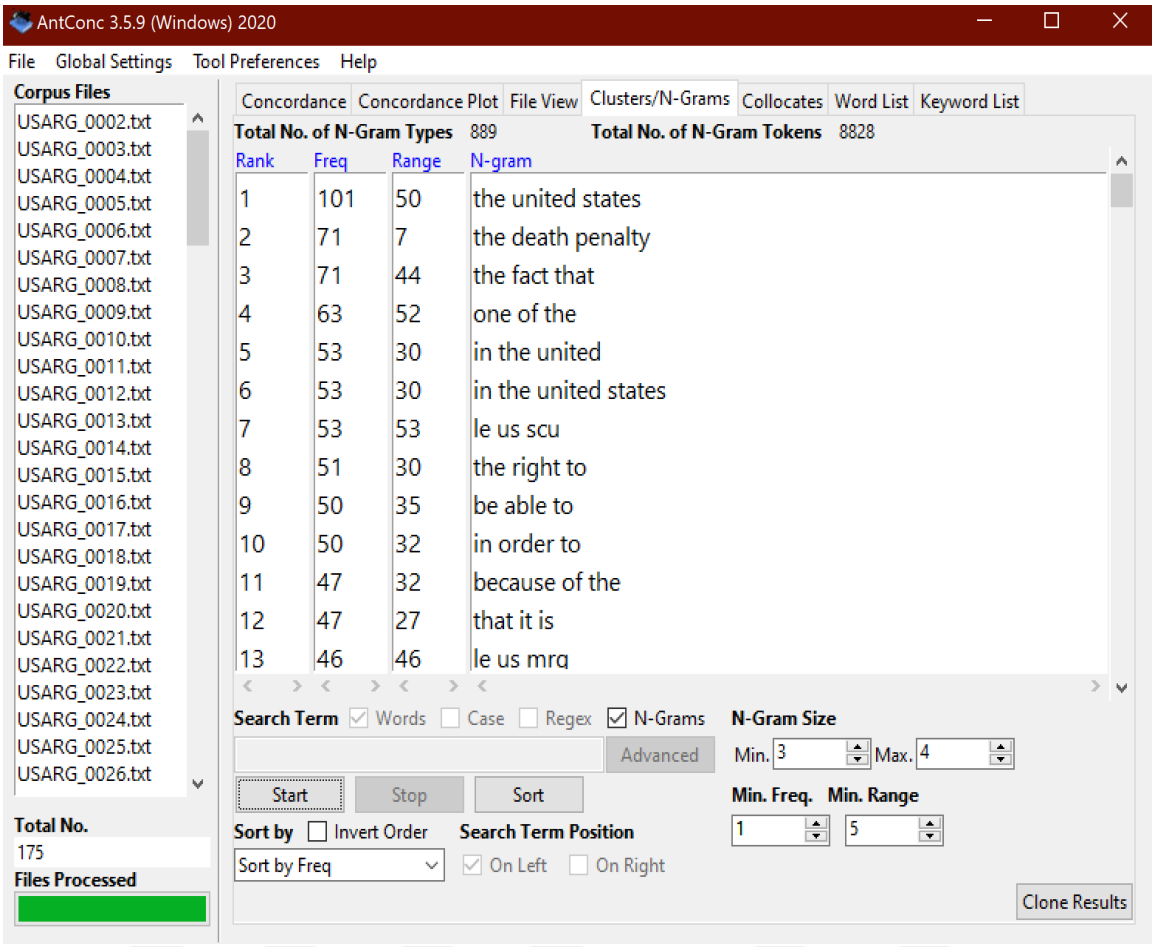


Figure 6. A Screenshot of AntConc Clusters/N-Grams Tool

Source: Anthony, 2020

Figure 4 presents the n-grams derived from the LOCNESS corpus in AntConC software, adhering to the aforementioned criteria. Referring to this analysis, we can observe the n-gram list in its entirety. Furthermore, to gain further insights into the usage and contextual occurrences of each specific n-gram, we can simply click on the desired item. This leads us to the concordance tool, where the surrounding contexts of the selected n-gram can be examined in detail.

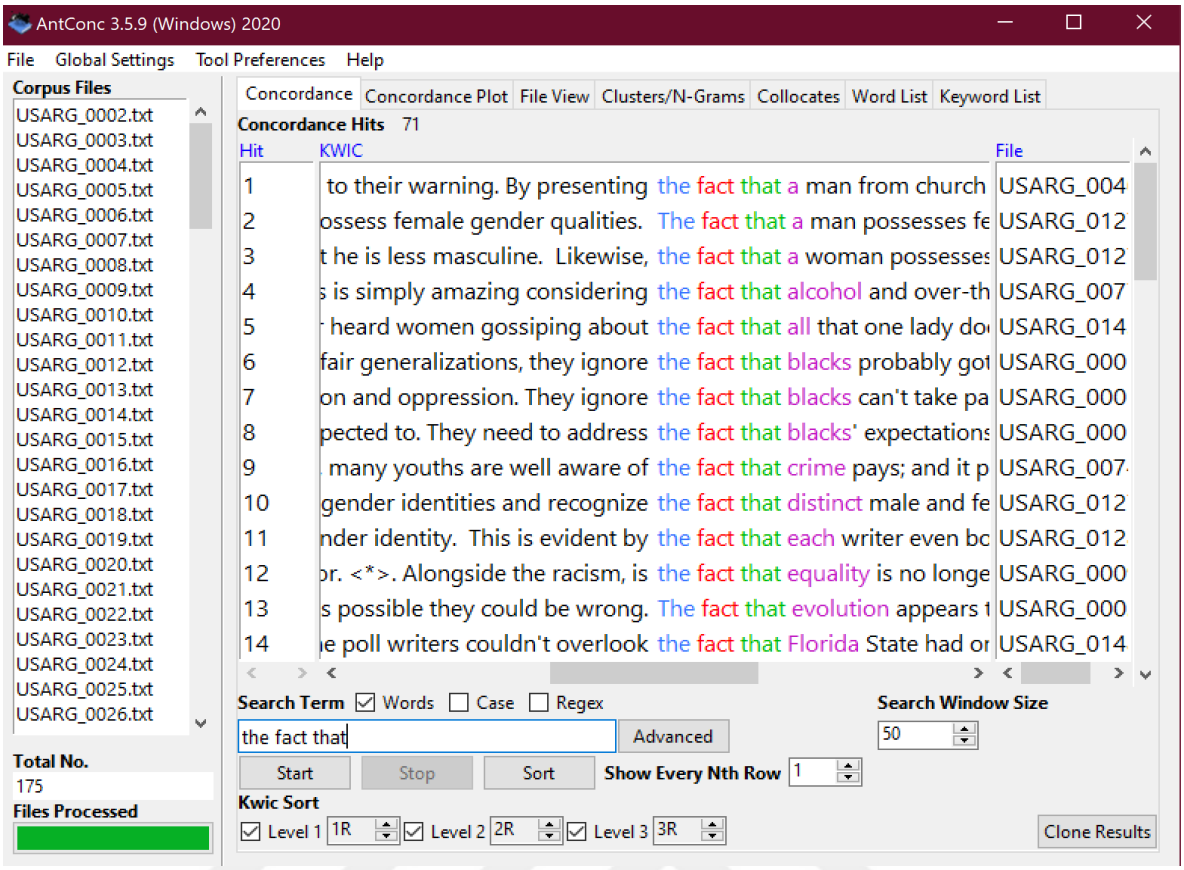


Figure 7. A Screenshot of N-grams in Context in AntConc

Source: Anthony, 2020

With the concordance tool, any words or phrases may be observed in their contexts. It is possible to see what kinds of items surround the words or phrases being examined. The occurrences along with their several examples may also be seen by this tool (Anthony, 2020).

In our study, all the text files were loaded to the AntConc. Clusters/N-Grams Tool has been chosen. N-gram size has been set to minimum to 3 and maximum to 4. Minimum frequency has been set to 1 in order to see all the 3 and 4-word sequences and determine the cut-off point manually by looking at the patterns. Minimum range has been set to 3 as MWIs have to appear in at least three to five different texts (Biber & Barbieri, 2007). And the phrases have been sorted by the frequency.

3.5.5. Wmatrix Corpus Analysis and Comparison Tool

Wmatrix is a web-based software designed as a corpus analysis tool which provides an interface to the corpus annotation tools, standard corpus linguistic methodologies including frequency lists and concordances. Wmatrix was initially

developed by Paul Rayson in the REVERE project. Later, it was extended and applied to corpus linguistics during PhD work and is still being updated regularly. The main page of this analysis tool may be seen below.

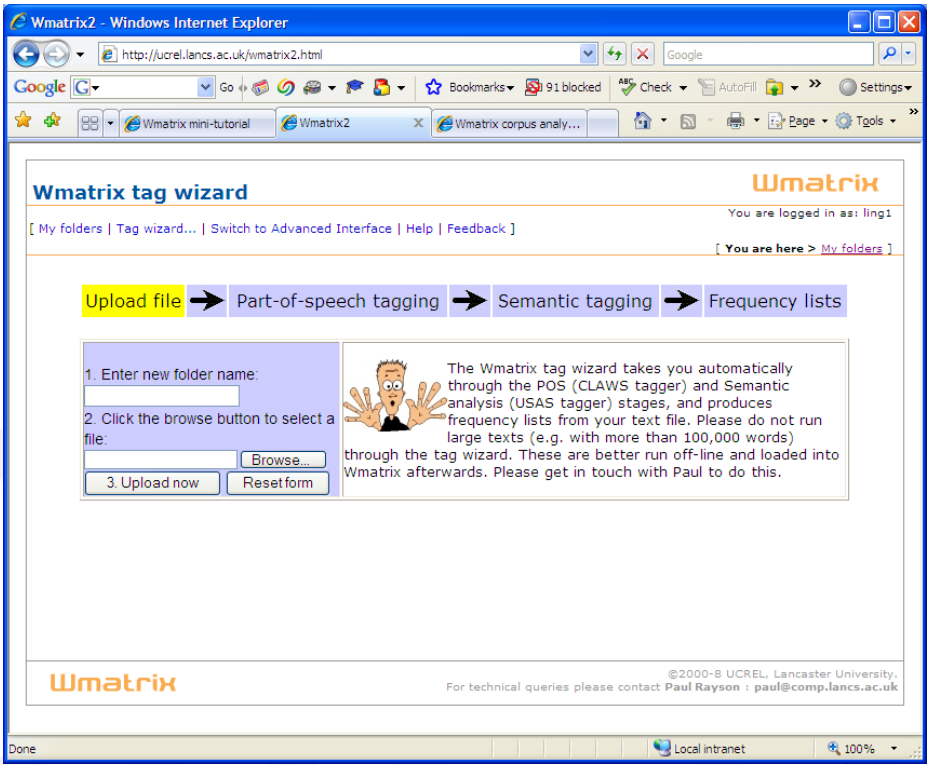


Figure 8. A Screenshot of Wmatrix main page
Source: Rayson, 2008

The users can run the tools on Wmatrix with a web browser like Chrome, Firefox, or Internet Explorer. The files may be uploaded and tagged according to the automatic tagging process by clicking on the tag wizard as can be seen from the figure. In addition, the web-based Wmatrix software (Rayson, 2008) permits the users to perform retrieval operations as well as annotating uploaded English texts with two levels of corpus annotation: part-of-speech and semantic field. This process can be conducted step by step starting with the 'load file without tagging' option.

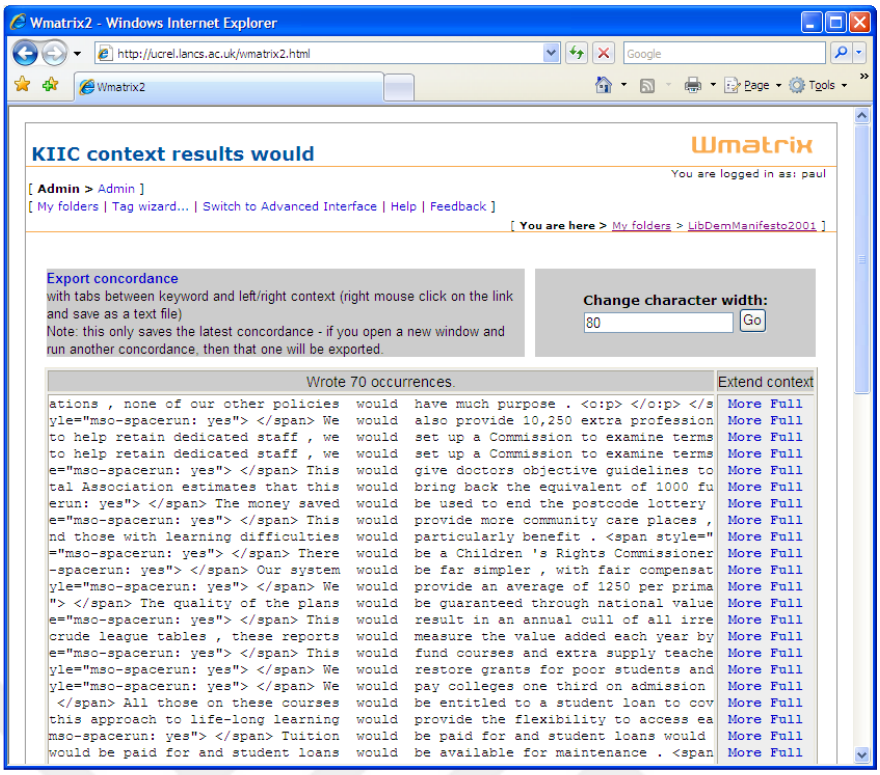


Figure 9. A Screenshot of Wmatrix key words in context

Source: Rayson, 2008

Key words in contexts can also be generated to see the most frequent items in the corpora. Frequency lists may be seen in the simple interface while POS tags and semantic tags are available in the advanced interface. The lists can be sorted either alphabetically or by frequency. The USAS tagger (Rayson et al. 2004b) which automatically assigns semantic fields (domains) to each word or multiword expression in the corpora has been generated. The figure below shows the multiword items and related tag sets for their semantic domains.

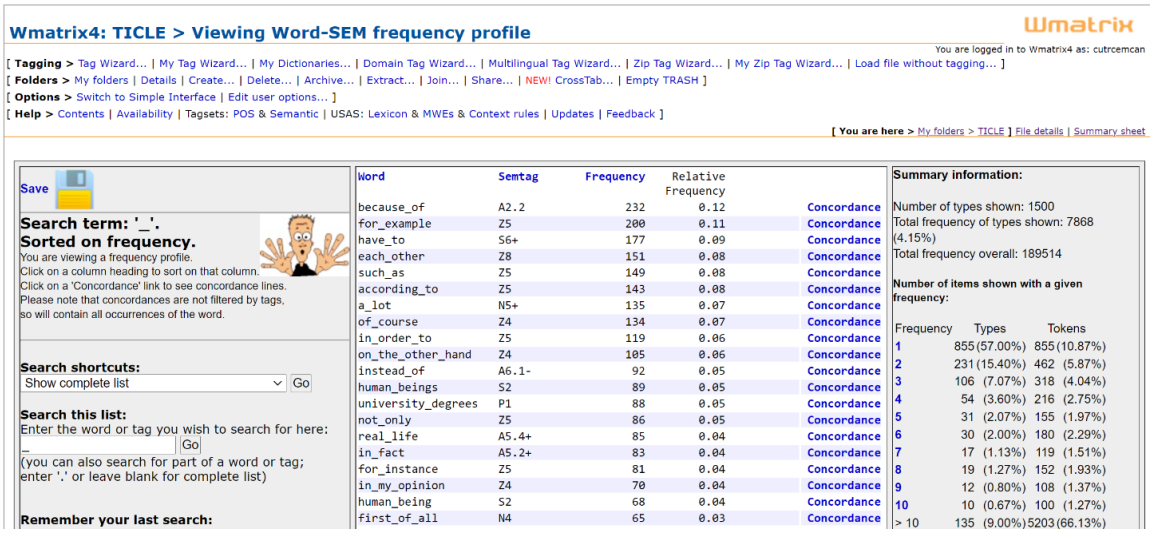


Figure 10. A Screenshot of Wmatrix word frequency profile

Source: Rayson, 2008

Wmatrix extends the keywords method to key grammatical categories as well as key semantic domains. From the frequency list view, 'concordance' can be reached. These can show the usual word-based concordance as well as all occurrences for words in one POS or semantic category. After we had the frequency lists for our both corpora, we chose the multiword expressions in USAS tagged file so that we could have the MWIs generated with their frequencies. After having the list, we listed the items in decreasing frequency. After that, the semantic domains for the most frequent MWIs have been generated manually from their tag sets one by one.

3.6. Structural and Functional Taxonomy of Lexical Bundles

Generated lexical bundles have been classified both structurally and functionally using a taxonomy developed by Biber, Conrad, & Cortes (2004, p. 381-396) which are summarized in Tables 11 and 12.

Table 11.

Structural Classification of Lexical Bundles

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

As can be seen from the classification, language formulas are classified according to three major structural features by Biber et al. (2004, p. 381) which are VP fragments, dependent clause fragments and noun phrase and prepositional phrase fragments. The original taxonomy (in p. 33) has been expanded by adding two more sub-categories by the researcher, first of which is *other adverbial clause fragments* added to the dependent clause fragments category, the second is *quantifier expressions* added to the noun phrases category. We have analyzed our formulas using this structural classification. The same formulas have also been categorized according to their functions that could be seen below:

Table 12.

Functional Classification of Lexical Bundles

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

Biber et al. (2004, p. 389-396) identified three primary discourse functions for language formulas in English: 1) stance expressions, 2) discourse organizers, and 3) referential expressions. Formulas serving similar functions have been grouped together based on their uses and meanings. This functional taxonomy has been used in the functional analysis of the lexical bundles in this dissertation.

CHAPTER IV

FINDINGS

4.1. Bird-eye View of the Findings

In the data analysis, a top-down approach has been adopted to gain insights about the data as well as findings by analyzing the overall picture first. In other words, in the data analysis process, we aimed to explore formulaic language in both TICLE and LOCNESS corpora. The number of words and LBs have been calculated by n-gram/clusters tool in AntConc analysis software. The total number of instances to be analyzed in both corpora has been determined.

After generating the sequences by frequency and range criteria, the manual selection was performed to eliminate the meaningless and same ones and decide the frequency cut-off point considering the total number of words in both corpora. While the cut-off point has been set as 10 for the TICLE corpus, the LOCNESS corpus has been cut off at 7. This has been decided to increase reliability and validity in both corpora by considering the results separately. And then, the frequency calculation has been made so that the frequency results may be compared to each other. The obtained data at the end was analyzed in its entirety to give a bird-eye view in terms of the number of words, the type and token ratio, and frequencies. The overall descriptive results of the lexical bundles may be seen in the Table 13 below.

Table 13.

The Bird-eye View of the Findings

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

As can be seen from the table, TICLE corpus consists of 280 texts with 199,532 words. LOCNESS corpus has 175 texts with 149,574 words. Total number of words used in the study is 349,106 while the total number of texts is 455. While Turkish EFL learners used 340 MWIs in their written productions, NS used 349 MWIs. Total number of MWIs used in the essays are 689. When looked at their percentages, only 0.17% of Turkish

learners' productions are considered bundles. As for the native speakers, 0.23% of their productions are lexical bundles. In total, 0.40% of these words consist of lexical bundles. The list of all of the bundles could be seen in Appendices A and B. In order to find out the statistical significance of these bundles in both corpora, log-likelihood analysis has been carried out.

4.1.1. Overlapping and Recurrent Bundles

While analyzing the extracted data, it was realized that some of the bundles were overlapping and recurrent. Recurrent ones have already been eliminated from the list. Overlapping bundles have been presented below. In the TICLE corpus, some of the 3-word bundles were part of the 4-word bundles. For this reason, the number of 4-word bundles has been subtracted from the 3-word bundles in order to eliminate the overlapping and have a more reliable token number. The overlapping bundles may be seen in the table below.

Table 14.

Overlapping Bundles in TICLE

MWIs	**Freq.**
a lot of	113
a lot of people	10
there is no	89
There is no need	17
but it is	29
But it is not	15
if there is	25
If there is a/no	23
around the world	12
All around the world	18
in front of	15
in front of the	11
important thing is	10
the important thing is	15
people who are	44
the people who are	21
as a result	47
as a result of	32
in addition to	12
İn addition to this/these	29
At the end	24
At the end of	13

As seen from the table, the number of overlapping bundles is 11. This means that only 11 3-word bundles out of 340 bundles were part of 4-word bundles. The number of 4-word bundles has been subtracted from the 3-word bundles in order to have a reliable count of the tokens. Also, each bundle in the table has been counted as a different type even though they have the same structure. The most frequent common bundle is *a lot of* with a frequency of 113. *There is no* follows as the second most common bundle with a frequency of 89. Also, third common bundle is *but it is* with a frequency of 29.

On the other hand, the LOCNESS corpus has also some overlapping bundles which are presented in Table 15 below.

Table 15.

Overlapping Bundles in LOCNESS

MWIs	**Freq.**
the death penalty	71
the death penalty is	17
One of the	63
One of the most	18
in the united states	53
the united states	48
This is a	31
This is a very	7
The end of	7
At the end of	9
The beginning of	7
At the beginning of	8

The number of overlapping bundles in the LOCNESS are 6. These 6 bundles are part of the 4-word bundles as can be seen. While counting, the number of 4-word bundles has also been subtracted from the 3-word bundles. Each of the bundles has been counted as a single type. The most frequent common bundle is *the death penalty* with a frequency of 71, while *the death penalty is* occurs 17 times. The second 3-word common bundle is *one of the* with a frequency of 63. And, *in the united states* follows as the third bundle occurring 53 times.

4.1.2. Type/Token Ratio Results

T/t ratio shows the lexical variety within a text. T/t ratio of lexical bundles represents the percentage of each bundle within all words in the corpora, that is, the number of bundles that falls into per 100 words. The T/t ratio of both corpora results may be seen in Table 16 below.

Table 16.

Type/token Ratio Results

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

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

As Table 16 indicates, TICLE corpus has 340 types and 8495 tokens while LOCNESS has 349 types and 4729 tokens. The breakdowns of each fragment are presented in the table. The total T/t ratio of the TICLE corpus is 4% in total. LOCNESS corpus has a T/t ratio of 7.3% in total. This means that 4 out of the Turkish learners' every 100 words are MWIs. Native speakers used 7.3 of their every 100 words as an LB. When we compare the two corpora considering their total number of words in both corpora, native speakers have produced more varied bundles according to their T/t ratio results.

4.2. Findings for the Research Question 1

RQ1: What are the most frequent 3- and 4-word lexical bundles found in argumentative essays of Turkish EFL learners in the Turkish International Corpus of Learner English (TICLE) and native speakers in the Louvain Corpus of Native English Essays (LOCNESS) corpora?

The first 100 most frequent expressions have been chosen in order to answer the first research question for the practical reasons. To be able to see the similarities and

differences between two corpora, the most frequent 100 LBs used in both TICLE and LOCNESS corpora are presented in Table 17 below. The common lexical bundles used in both corpora have been written bold and italic.

Table 17.

Top 100 Lexical Bundles in order of Frequency in both Corpora

TICLE		**LOCNESS**	
LBs	**Freq.**	**LBs**	**Freq.**
men and women	***126***	the death penalty	71
a lot of	***123***	***the fact that***	***71***
it is not	***120***	one of the	63
in order to	***119***	in the united states	53
on the other hand	***119***	***the right to***	***51***
they do not	***119***	***in order to***	***50***
the most important	***109***	the united states	48
in the world	***108***	***because of the***	***47***
there is no	***106***	***that it is***	***47***
most of the	***98***	the use of	45
there is a	***86***	***it is a***	***40***
the real world	83	as well as	40
as a result	79	***it is not***	***39***
it is a	**76**	in public schools	36
there are some	72	to be able to	35
do not have	***71***	***there is no***	***35***
in my opinion	70	***that they are***	***33***
it is the	***69***	***there is a***	***33***
the people who	67	***this is a***	***31***
first of all	65	***part of the***	***31***
people who are	65	***should not be+V3***	***30***
they are not	***65***	***a lot of***	***29***
that it is	***64***	***all of the***	***29***
there are many	***62***	invention of the	29
i want to	59	***men and women***	***29***
because of the	***58***	due to the	28
in our country	57	the idea of	28
cannot be	53	***to be a***	***28***
they can not	53	***they do not***	***27***
to have a	***51***	he or she	26
women and men	51	out of the	26
day by day	50	***the number of***	***26***
should not be+V3	***33***	his or her	24

in the society	49	***there are many***	***24***
that they are	***49***	have to be	23
is not a	***47***	in favor of	23
is the most	45	**in the past**	**23**
but it is	44	is not the	23
prepare students for	44	the amount of	23
i believe that	43	this type of	23
that there is	43	to be the	23
the right to	***43***	as long as	22
university degrees are	43	***is not a***	***22***
according to the	***41***	it would be	22
they want to	***41***	***on the other hand***	***22***
in addition to	41	the case of	22
in the past	***41***	the end of	22
the real life	41	***to have a***	***22***
is one of the	***40***	is a very	21
the help of	40	***it is the***	***21***
to sum up	40	***you do not***	***20***
i do not	39	a part of	20
to be a	***39***	***according to the***	***20***
if there is	38	i feel that	20
point of view	37	in the case	20
in the future	***36***	***they want to***	***20***
they have to	36	seem(s) to be	19
look at the	***35***	because it is	19
some of them	35	***because they are***	***19***
this is a	***34***	***in the world***	***19***
and they are	33	***most of the***	***19***
if you have	33	some of the	19
it should be	51	the root of	19
there are a lot	33	way of life	19
all over the world	32	whether or not	19
as a result of	***32***	would not be	19
it will be	32	one of the most	18
part of the	***32***	the effects of	18
we can not	32	the majority of	18
and do not	31	***the most important***	***18***
because they are	***31***	many of the	17
have the same	31	the death penalty is	17
it is very	31	to make a	17
you can not	31	***is one of the***	***16***
the fact that	***31***	is that the	16
there will be	31	it does not	16

we can see	30	it has been	16
one of them	30	such as the	16
that there are	30	the lack of	16
this kind of	30	***they are not***	***16***
we look at	30	***this is not***	***16***
all of the	***45***	this is the	16
all of them	16	to say that	16
you do not	***29***	***as a result of***	***15***
do not want	***29***	be allowed to	15
i think that	29	***do not have***	***15***
in our society	29	***look at the***	***15***
you want to	29	the beginning of	15
most of them	29	the end of the	15
the importance of	29	the opposite sex	15
this is not	***29***	there has been	15
with each other	29	to deal with	15
and it is	27	up to the	15
as it is	27	would be a	15
because of this	27	an example of	14
for example in	27	it can be+V3	14
in the family	27	***do not want***	***14***
know how to	27	***in the future***	***14***
That they can	27	the issue of	14
The number of	***27***	the question of	14

Comparing the LBs helps us understand the similarities and differences between the NS and EFL learners. While comparing, it is necessary to see which items have been used commonly. As can be seen from the table, 40 common bundles exist in the most frequent 100 bundles of both corpora. These bundles are *men and women, a lot of, it is not, in order to, on the other hand, they do not, the most important, in the world, there is no, most of the, there is a, it is a, do not have, it is the, they are not, that it is, there are many, because of the, to have a, should not be+V3, that they are, is not a, the right to, according to the, they want to, in the past, is one of the, to be a, in the future, look at the, this is a, as a result of, part of the, because they are, the fact that, all of the, you do not, do not want, this is not,* and *the number of.* These bundles are used more frequently in TICLE corpus relative to LOCNESS corpus.

As for the structures of these common expressions, while 19 of these expressions belong to the *verb phrase* fragments, the others are included in the *noun phrase and prepositional phrase* fragments. 14 of these bundles are *personal pronouns+VP*

fragments such as *they are not, that it is, there is no*…etc. Nine of the expressions are included in the *prepositional phrase* expressions category. Examples include *in order to, the number of, part of the, in the past, in the future,* and *in the world.* Five of the expressions belong to the *VP with non-passive verb* category like *do not want, look at the, do not have…*etc. There are 3 *other noun phrase expressions* which are *men and women, according to the,* and *the right to.* Another sub-category which has been added to the noun phrase and PP fragments by the researcher is the *quantifier expressions* which includes the expressions like *a lot of, most of the* and *all of the.* Also, 3 of the expressions were *of-phrase* expressions of which examples include *as a result of, part of the* and *the number of.* There is only one comparative expression that is *the most important.* And lastly, two of the expressions belong to the *that-clause* fragments which are *that it is* and *the fact that.*

When the frequency of these bundles compared, the most frequently used lexical bundle by the Turkish EFL learners is *men and women* (f: 126) whereas the most frequently used lexical bundle by the NS is *the death penalty* (f: 71). The most frequent expressions in TICLE corpus are *a lot of* (f: 123), *it is not* (f: 120), *in order to* (f: 119), *on the other hand* (f: 119), *they do not* (f: 119), *the most important* (f: 109) respectively. Whereas in LOCNESS corpus, the most frequent expressions are *the fact that* (f: 71), *the right to* (f: 51), *in order to* (f: 50), *because of the* (f: 47), *that it is* (f: 47). These 10 most frequent bundles (1147 tokens) constitute 13% of the whole tokens in TICLE. 10 most frequent bundles in LOCNESS occurred 546 times amounting 11% of all the tokens. As can be seen, there exists an immense difference between two corpora regarding the frequency of the common expressions.

Besides, by examining the use of pronouns in all lexical expressions, it was also revealed that Turkish learners have used singular and plural personal pronouns more often than the NS. 62 singular and 72 plural pronouns have been used in total by Turkish EFL learners whereas 34 singular and 25 plural pronouns have been used by NS in total. Looking at their LL values (4.43+), we found that all pronouns have been overused by Turkish EFL learners except for the singular pronoun *I.*

4.3. Findings for the Research Question 2

RQ2. How are these lexical sequences found in TICLE and LOCNESS corpora classified based on their structural and functional characteristics?

Firstly, structural and functional classifications of both TICLE and LOCNESS have been analyzed in detail. After that, comparisons of the two corpora have been presented.

4.3.1. Structural Classification of Lexical Bundles in TICLE

This research question investigates the structures of the extracted and classified LBs according to Biber et al.'s (2004) taxonomy as stated in the previous chapter. However, some of the bundles could not be categorized according to the existing categories. That is why, two new sub-categories have been added to the original categorization. The first one is *other adverbial clause fragments* within the dependent clause fragments such as *since they are, but it is not, as I said*. The latter is *quantifier expressions* such as *a lot of, some of them, most of the people* added to the noun phrase and prepositional phrase category. Firstly, the overall table including frequencies, types and tokens of the bundles are presented. After that, the detailed structural classifications are given.

Figure 11 show the distribution of the structural types in general.

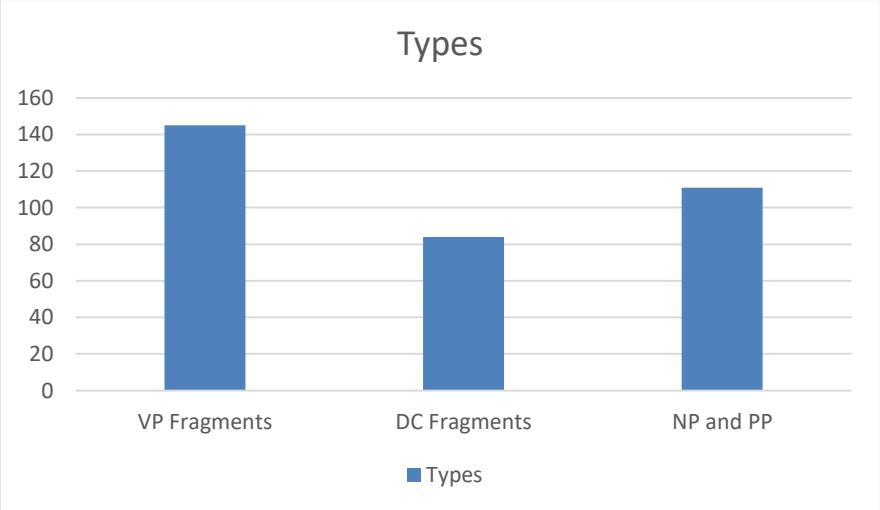


Figure 11. Distribution of structural types in TICLE

Figure shows the structural types in TICLE. While the most frequent category seems to be the VP fragments, second most frequent category belongs to NP and PP fragments. The least used bundles by Turkish EFL learners are dependent clause fragments. This reveals that Turkish EFL learners mostly prefer using clausal bundles.

As for the number of the tokens of these lexical phrases, they are consistent with

the number of types. *Personal pronouns+VP fragments* are the most used structures accounting a third of all the tokens in the structural classification. The second highest percentage belongs to *prepositional phrase expressions* with the percentage of 18%. The detailed summary of the results is shown in Table 18 below.

Table 18.
Structural Classification of Lexical Bundles in TICLE

STRUCTURE	**TYPES**	**%**	**TOKENS**	**%**
1. Verb Phrase Fragments				
a. 1st/2nd/3rd person pronoun + VP fragment	115	33.8	2672	33
b. Discourse marker + VP fragment		-		-
c. Verb phrase with non-passive verb	19	5.5	588	6
d. Verb phrase with passive verb	10	2.9	194	2
e. Yes/no question fragments	1	0.2	12	1
f. WH-question fragments		-		-
2. Dependent Clause Fragment				
a. 1st/2nd/3rd person pronoun + dependent clause fragment	15	4.4	319	3
b. WH-clause fragments	9	2.6	140	1
c. *If*-clause fragments	14	4.11	239	4
d. (copula be+adjective/noun phrase) *To*-clause fragments	13	3.2	407	3
e. *That*-clause fragments	10	2.9	280	3
f. Other adverbial clause fragments	23	6.7	470	7
3. Noun Phrase and Prepositional Phrase				
a. Noun phrase with of-phrase fragment	14	4.11	394	4
b. Noun phrase with another post-modifier fragment	5	1.4	143	1
c. Other noun phrase expressions	22	6.4	763	8
d. Prepositional phrase expressions	56	17	1363	17
e. Comparative expressions	4	1.1	166	2
f. Quantifier expressions	10	2.9	345	4
TOTAL	**340**	**100**	**8495**	**100**

Table 18 displays the structural classification of MWIs with the corresponding type and token frequencies. As can be seen from the table, there exist 340 types and 8495 token in total of structural classifications of TICLE.

The first category is *the verb phrase fragments* which are the most frequent structures in TICLE. 145 bundles of 340 types belong to this category which has the widest variety among all other lexical-bundle structures. VP fragments are the most common structure used with 42.4% usage by the learners. When looked at the verb phrases in detail, *personal pronouns + VP Fragments* category account for a quarter of all structural classification. Within the VP fragments, Turkish EFL learners have not used any discourse markers such as *I mean you, you know it was and I mean I don't* and wh-question fragments including *what do you think, how many of you, what does that mean…etc.* Also, yes-no question fragment was only used once in TICLE.

Second category is dependent clause fragments with 84 types in total. This category constitutes only 24% of the whole structural classifications. *Other adverbial clause fragments* are the most common structure with 23 different types. *Personal pronouns + dependent clause fragment* follows as the second most common structure. *Wh-* and *that-clause* fragments have both been used as the least common structures with 9 and 10 types respectively.

The last category is NP and PP structures with 111 different bundles. This category is the second most common category in structural categorization with a percentage of 33%. The prepositional phrase expressions have been used with 56 different types, having the highest number of tokens within this third section. The second most common types were used as *other noun phrase expression*s with 763 tokens. And the least used fragment is comparative expressions in this group with only 4 types.

4.3.1.1. Verb Phrase Fragments in TICLE

The first fragments of structural classifications are verb phrase fragments which consist of *personal pronouns + VP fragment, discourse marker + VP fragment, VP with non-passive verb, VP with passive verb, yes/no question fragments, wh-question fragments*. However, only four of these categories have been produced in TICLE. While "*yes/no question fragments*" has only been used once in the whole corpus, the *"discourse marker + VP fragment"* and *"WH-question fragments"* have never been produced at all. All types of VP fragments with their tokens in brackets have been given in Table 19 below.

Table 19.

Verb Phrase Fragments in TICLE

a. 1st/2nd/3rd person pronoun + VP fragment	Students/universities/people/I/you/they/we/it do(es) not (f: 214), they/we/people are not (f: 90), I/they/you/he want(s) to (f: 139), university degrees are (f: 43), I think the (f: 20), I think it is (f: 12), men and women have(f: 15), it is used (f: 10), one of them is (f: 13), they try to (f: 21), animal testing is (f: 23), money is the (f: 24), they had to (f: 10), they may have (f: 10), they have got (f: 11), they need to (f: 12), there have been (f: 15), it has been (f: 19), sex equality is (f: 19), they start to (f: 22), you have money (f: 18), you have a (f: 12), we look at (f: 30), they do not have (f: 13), they do not want (f: 11), they have a/some (f: 35), we know that (f: 16), we all know (f: 12), I/we/they have to (f: 72), important thing is (f: 10), the important thing is (f: 15), there are some/many/a lot/lots of/many people/ also/so many (f: 215), there is a/an (f: 108), there is no (f: 89), there is no need (f: 17), there is not (f: 26), there is nothing (f: 20), there/women/euthanasia/it/they should be (f: 97), they/it should not be (f: 25), they should not (f: 11), the teachers should (f: 10), I/they/you/we/he/women cannot (f: 194), you/we can see (f: 44), we can give (f: 11), it/this/they can be (f: 60), I/we can say that (f: 27), you/they can do (f: 33), it is true/a/an/the/possible/impossible/very/also/better/fact/necessary/so (f: 292), it is not (f: 99), it is not a (f: 10), it is not enough (f: 11), it is obvious that/true that (f: 20), this is a/the (f: 60), this is not (f: 29), that is the (f: 16), I/it/they/there will be (f: 96), they will not (f: 13), they will have (f: 10), euthanasia is the (f: 15), we do not have (f: 14), they do not know (f: 11), it can be+adj (f: 13)
b. Discourse marker + VP fragment	-
c. Verb phrase with non-passive verb	do not have/want/prepare/know (f: 140), is not a/the/enough (f: 92), is one of the (f: 40), is the most (f: 24), is the most important (f: 21), have the same (f: 31), look at the(f: 33), know how to (f: 27), get rid of (f: 21), prepare students for (f: 44), have the right (f: 33), cannot be +adj (f: 25), should not be+adj (f: 17), will be a (f: 20)
d. Verb phrase with passive verb	should be given (f: 17), be used in (f: 10), to be used (f: 10), to be killed (f: 12), can be done (f: 17), can be used (f: 13), cannot be+V3 (f: 52), should not be+V3 (f: 33), thought to be (f: 19), going to be (f: 11)
e. Yes/no question fragments	do you think (f: 12)
f. WH-question fragments	-

As can be inferred from the table above, VP fragments are the most frequent structures in TICLE, accounting for 42% of all LBs in the corpus. Personal pronouns+ VP fragments comprise the 34% of all the VP fragments. This category feature 145 different types, the widest variety among all other lexical bundles structures. The tokens amount to 3466. Type and token ratio is 4.1% in verb phrase fragments. Even though they have the widest variety of types, most of their structure is within *personal pronouns+ VP fragments* category. These structures feature more variations than the other 2 categories.

Being the most common type of lexical bundles, *personal pronouns+VP* fragments consist of a subject pronoun followed by a verb phrase. This category holds expressions with all kinds of pronouns such as *I, you, we, they, s/he, it*. However, the highest number of types start with the "it is+noun/adj phrase" structure with a frequency of 292. Some of the examples may be seen below:

(1) **it is a** better invention than telephone because you [TRCU1016]
(2) As you see, **it is impossible** to live without money. [TRCU1056]
(3) not know the content of it or why **it is so** important to me to be a [TRCU1066]
(4) Nowadays, **it is very** popular to find friends by the [TRCU1015]
(5) In my opinion, **it is better** to die than to suffer from [TRCU1161]

The second highest number of types belongs to the "existential *there*-constructions" with 458 types in total. While "there is" structure has 243 types in the corpus, "there are" structure bear 215 types. The examples in our study include *there are some/many/a lot/lots of/many people/ also/so many* (f: 215), *there is a/an* (f: 108), *there is no* (f: 89), *there is no need* (f: 17), *there is not* (f: 26), *there is nothing* (f: 20). The examples of this structure produced by Turkish EFL learners are given below:

(6) **There are some** countries that have abolished this form [TRCU1104]
(7) we can realize that **there are lots of** things became a part of [TRCU1018]
(8) So, **there is no need** for male dominance. [TRCU1115]
(9) **there is a** balance between school community and individual [TRKE2011]

Another significant number of types seems to belong to the ability modal "can". While the affirmative sentences hold 148 types, it has been used negatively with 194 different types. Some examples are as follows:

(10) **Women cannot** make a stand against men. [TRKE2025]

(11) **We cannot** calculate how much they have been learned. [TRCU1063]

(12) In conclusion, **I can say that** we cannot limit education of [TRME3026]

In addition, there are also some other modals used within this category such as *I/we/they have to, they should not, I/it/they/there will be, they will have* of which examples are as such:

(13) Perhaps I find the answer when **I will be** a teacher. [TRCU1004]

(14) is not an help to their friends and **they should not** consider it as a way to [TRKE2026]

(15) Consequently, it seems that **we have to** choose, with nuclear energy, we can [TRCU1077]

Within the VP fragments, large number of the phrases report negative states which can be seen in the examples of *students/universities/people/I/you/they/we/it do(es) not* (f: 214), *they/we/people are not* (f: 90), *they/it should not be* (f: 25), *they should not* (f: 11), *I/they/you/we/he/women cannot* (f: 194), it is not (f: 99), it is not a (f: 10), it is not enough (f: 11), this is not (f: 29), they do not have (f: 13), we do not have (f: 14)…etc. These negative structures constitute 22.6% of the personal pronouns+VP fragments subcategory.

The third category includes *verb phrases with non-passive fragments* which are phrases plus a part of a following complement phrase or clause. There exist 20 types in total with 588 tokens. Some examples of this category can be seen below.

(16) There are still lots of people who **do not know** anything about computers who do not… [TRCU1012]

(17) When we **look at the** improvements in medicine, we see that [TRCU1140]

(18) Divorce **is the most** common subject that people is talking… [TRCU1168]

The fourth category has *"verb phrase with passive verb"* fragments which have 10 types in total. All of the bundles in this subcategory have been structured with modals by Turkish EFL learners. Passive usages of "can" and "should" modals have been separated from "it can/should be+adj (13)" fragments. Some passive structures have been

given below.

(19) To all human life **should be given** equal protection under the law. [TRCU1091]

(20) No one deserves **to be killed** whatever the crime is because [TRCU1105]

(21) Cheating **cannot be annihilated** completely. [TRCU1162]

(22) Euthanasia **should not be allowed** since it has many objections [TRCU1140]

5th category belongs to *"yes/no question fragments"* which begin with either an auxiliary verb or a modal verb, followed by a subject pronoun. Some of the fragments include *are you going to, do you want to, does that make sense.* This type has solely been used as a one single type in the whole TICLE. That usage is "*do you think*" with 17 types which examples may be seen below.

(23) **Do you think** that everything is equal for us? [TRCU1137]

(24) **Do you think** that men and women are treated [TRKE2070]

As for the *"discourse marker + VP fragment"* such as *I mean you, you know it was and I mean I don't* and *"WH-question fragments"* including *what do you think, how many of you, what does that mean…etc*. have not been produced by Turkish EFL learners at all, which is another striking finding in the corpus.

Lastly, Turkish EFL learners have used many of the structures with modals in VP fragments. This is particularly obvious in phrases with passive verb subcategory. 51 of the types include modals, constituting 35.1% of all verb phrase fragments. The next section will analyze the dependent clause fragments in detail.

4.3.1.2. Dependent Clause Fragments in TICLE

The second structural category in Biber et al.'s (2004) structural classifications of lexical bundles are dependent clause fragments. Sub-categories of this fragment are 1st/2nd/3rd person pronoun + dependent clause fragment, wh-clause fragments, if-clause fragments, (copula be+adjective/noun phrase) to-clause fragments, that-clause fragments. However, in the analysis process, some of the bundles could not be categorized into any of the existing slots. That is why, another category has been propounded to the existing classification as *other adverbial clause fragments* by the researcher, which is presented in the following table.

Table 20.

Dependent Clause Fragments in TICLE

a. 1st/2nd/3rd person pronoun + dependent clause fragment	I/you/we/they/people think that (f: 92), I believe that (f: 43), we see that (f: 23), people/students/women who are (f: 69), the people who are (f: 21), person who is (f: 17), people who have (f: 22), it means that (f: 19), this means that (f: 13)
b. WH-clause fragments	when we look at (f: 21), when we think (f: 10), when they are (f: 21), when it is (f: 17), when they want (f: 11), whatever they want (f: 12), what they have (f: 10), how to use (f: 21), what to do (f: 17)
c. *If*-clause fragments	if you/they have (f: 47), if we think (f: 14), if they are (f: 21), if they do (f: 17), if you are (f: 26), if it is (f: 15), if we look (f: 12), if they do not (f: 12), if they do(f: 17), if there are (f: 10), if there is no (f: 10), if there is (f: 25), if there is a (f: 13)
d. *To*-clause fragments (copula be+adjective/noun phrase)	in order to (f: 119), in order not to (f: 12), to find a (f: 22), to make the (f: 21), to be able to (f: 26), to sum up (f: 40), to be a (f: 39), to live in (f: 22), to say that (f: 20), to go to (f: 24), to have a (f: 51), would like to (f: 11), have to be (f: 20),
e. *That*-clause fragments	that it is (f: 52), that it is a (f: 12), that there is/are (f: 73), that they are (f: 49), that they can (f: 27), that women are (f: 22), that they have (f: 21), that we have (f: 12), that you are (f: 12)
f. Other adverbial clause fragments	and they are (f: 33), and it is (f: 27), and do not (f: 31), and they can (f: 24), because it is (f: 22), however it is (f: 11), in fact it is (f: 12), since they are (f: 13), although it is (f: 15), because they are (f: 31), so it is (f: 13), so there is (f: 13), so they can (f: 13), but it is (f: 29), but it is not (f: 15), but there is (f: 20), but they are not (f: 11), but they are (f: 8), as you see (f: 20), as I mentioned (f: 14), as I said (f: 13), as it is (f: 13), as we know (f: 13)

All types of dependent clause fragments with their token counts in the corresponding brackets are listed in the table. There exist 84 types with 1855 tokens in total which concludes to a T/T ratio of 4.3%. This fragments is the least common category among the structural classifications, constituting only 24% of the whole classification.

The first category is personal pronouns + dependent clause fragments which starts with any pronouns followed by dependent phrases such as *I believe that, we see that, people/students/women who are.* This category is the second most common category of

dependent clause fragments and most of the LBs include *people* or collocate with *people*. Some examples of this category are as follows.

(25) Some other **people think that** divorce is inevitable because the member [TRCU1163]
(26) Today there are a lot of successful **women who are** leading whole companies and even whole [TRCU1035]
(27) In most films **we see that** some people fight for a big [TRCU1121]

The second category is *WH-clause fragments* which have two types; verb phrases followed by a *wh*-clause or bundles with a *wh*-clause fragments. Our data only consisted of the second type, of which the examples may be seen below.

(28) **When we look at** the role of the universities, [TRCU1063]
(29) my point of view people are guilty for **what they have** done and also what forces them [TRCU1041]
(30) at that moment they can do **whatever they want** because of having money. [TRCU1133]

The third category belongs to the *if-clause fragments* with 14 different types and 239 tokens in total, becoming the third most common category of the dependent clause fragments. Examples of this category may be seen below.

(31) **If you have** a computer and also internet, it [TRCU1015]
(32) **If there is no** solution for the patient except [TRCU1084]
(33) **If they are** important subjects for all the people. [TRKE2012]

Fourth category is *(copula be+adjective/noun phrase) To-clause fragments* which are of three types: verb phrases followed by a to-clause, bundles that represent the beginning of a to-clause and predicative adjective+to-clause. *To sum up* and *to be a* have been used prominently in TICLE. Some of the examples are presented as follows.

(34) **To sum up,** computer is the most important invention [TRCU1015]
(35) Although this seems **to be a** bad situation in a social life [TRCU1154]
(36) tongues when I turn on the TV or **to be able to** buy and read publications that

[TRCU1126]

Another category is *that*-clause category which has two main types: those including main clause verb, and those comprising only the *that*-clause. Our data consisted only the second main category, whose examples may be seen below.

(37) This would mean **that it is** acceptable for an individual to die [TRCU1173]
(38) I do not think **that there is** a full sexual equality and sex [TRKE2003]
(39) Men always think **that they are** in higher positions than women. [TRKE2033]

The last category is the other adverbial clause fragments with 22 different types, becoming the most common phrases in dependent clause fragments. This category has been added to Biber et. al.'s (2004) structural categorization by the researchers since some of the bundles could not fit in any of the existing classifications. These fragments begin with adverbial clauses of *as, but, because, so, although* and *however*. Some of the examples are as such:

(40) **As you see** from the examples education in university [TRCU1072]
(41) the people is not alone in the world **and they are** compelled to the nature. [TRKE2003]
(42) carry the responsibility on their shoulders **but it is** a fact that in the past, [TRKE2057]

In the next section, we will analyze the noun phrase and prepositional phrase fragments in detail.

4.3.1.3. Noun Phrase and Prepositional Phrase Fragments in TICLE

The third and the last structural category according to Biber et al.'s (2004) structural classifications of lexical bundles are noun phrase and prepositional phrase fragments. Sub-categories of this fragment are noun phrase with of-phrase fragment, noun phrase with another post-modifier fragment, other noun phrase expressions, prepositional phrase expressions, comparative expressions, quantifier expressions. We added a category, quantifier expressions, to NP and PP fragments so that some of the bundles such

as *a lot of, all of them, most of them, most of the people* could be categorized. Table 21 represents the fragments with all the types as well as their token counts in brackets.

Table 21.

Noun Phrase and Prepositional Phrase Fragments in TICLE

a. Noun phrase with of-phrase fragment	first of all (f: 65), because of the (f: 58), this kind of (f: 30), part of the (f: 32), the number of (f: 27), aware of the (f: 23), because of this (f: 27), the beginning of (f: 22), the rest of (f: 17), every part of (f: 21), the result of (f: 20), the use of (f: 22), the root of (f: 20), the value of (f: 22)
b. Noun phrase with another post-modifier fragment	the people who (f: 67), many people who (f: 16), a person who (f: 26), some people who (f: 14), the one who (f: 20)
c. Other noun phrase expressions	men and women (f: 102), both men and women (f: 24), the real world (f: 83), the real life (f: 41), point of view (f: 37), women and men (f: 51), the other people (f: 27), his or her (f: 24), the students to (f: 27), and so on (f: 20), the fact that (f: 31), a chance to (f: 25), the other one (f: 17), and the other (f: 21), for example the (f: 23), the right to (f: 43), according to the (f: 41), according to me (f: 14), day by day (f: 50), to each other (f: 20), as a conclusion (f: 22), as well as (f: 20)
d. Prepositional phrase expressions	in addition to (f: 29), in addition to this (f: 17), in addition to these (f: 12), in the past (f: 41), in real life (f: 16), in the exams (f: 28), in the family (f: 27), at this point (f: 21), in spite of (f: 20), on the other hand (f: 119), in the future (f: 36), in our society (f: 29), in front of (f: 15), in front of the (f: 11), in the society (f: 49), in my opinion (f: 70), all over the world (f: 32), in the world (f: 108), around the world (f: 12), all around the world (f: 18), in our country (f: 57), at the end (f: 24), at the same time (f: 24), for example in (f: 27), with each other (f: 29), the importance of (f: 29), in this century (f: 12), in other words (f: 21), in this way (f: 10), in the house (f: 18), in the universities (f: 15), in the real world (f: 16), in this situation (f: 16), at the beginning (f: 13), in the end (f: 13), in our lives (f: 14), in a society (f: 11), in the field (f: 12), in the life (f: 12), for a long time (f: 16), for real world (f: 16), for the people (f: 23), for the students (f: 23), for the sake of (f: 16), related to the (f: 20), for this reason (f: 21), a kind of (f: 20), the help of (f: 40), at the end of (f: 13), as a result of (f: 32), in terms of (f: 18)
e. Comparative expressions	as much as (f: 20), more important than (f: 15), more and more (f: 22), the most important (f: 109)
f. Quantifier expressions	a lot of (f: 113), a lot of people (f: 10), all of them (f: 32), one of them (f: 17), most of them (f: 29), some of them (f: 35), some of the (f: 23), all of the (f: 45), most of the people (f: 21), all the people (f: 20)

Noun phrase and prepositional phrases have 111 types with 3174 tokens in total, with a T/T ratio of 3.4%. This category is the second most common category within the structural classifications, constituting 33% of all the bundles.

To start with, the first category is noun phrase with *of*-phrase fragments, conveying a range of meanings like physical identification of a place, size and amount. Some of the examples are as follows:

(43) Firstly in every country the biggest **part of the** management is under control of men [TRKE2009]
(44) which are known as developing countries, **the number of** the educated women is quite low. [TRCU1135]

In addition, other bundles identify various abstract qualities such as:

(45) human health could not have been overcome without **the use of** animal testing. [TRCU1141]
(46) Then this decreases the value of some universities degrees. [TRME3015]

Some of the *of*-phrase bundles within this category indicate simple existence or presence such as *the presence of, the existence of*. Some others describe processes or events lasting over a period of time like *the development of, the course of the*...etc. TICLE did not have any bundles belonging to these two categories.

The second category is *noun phrase with another post-modifier fragment* which are composed of a noun phrase followed by a post-modifier other than an *of*-phrase. These have two major types: a noun phrase with a post-nominal clause fragment, and a noun phrase with a prepositional phrase fragment. TICLE included only the first types, of which examples are presented below.

(47) I believe that most of **the people who** apply to euthounasia can not have [TRCU1076]
(48) Also, there are **many people who** are against to the abortion because [TRCU1094]
(49) Of course we can do it by being **the one who** choose the programmes. [TRCU1029]

The third category belongs to the *other noun phrase expressions* with 23 different types and 763 tokens in total, becoming the second most common category of the NP and PP fragments. These expressions do not fit into any of the existing categories. Some of the examples can be seen as follows:

(50) In some societies, **men and women** are not looked at the same [TRCU1120]
(51) If students are kept away from **the real world**, and just look through that glass, [TRCU1067]
(52) **According to the** article on the internet; in June, 1987, [TRCU1166]

Fourth category in this section is *prepositional phrase expressions* with 56 types and 1472 tokens. This category is the most frequent category within the noun phrase fragments. PPs with a preposition head have two types which are prepositional phrases or prepositional phrase fragments with an *of* following the noun, and other prepositional phrases or prepositional phrase fragments. Some of the examples from TICLE are as follow:

(53) think that money is the most important thing **in the world** and they try to do almost [TRCU1146]
(54) it should make initiative **in the society** and we should be graduated from [TRCU1147]
(55) very common, but when we look **at the end of** the 20th century, we can [TRCU1015]

Another category is *comparative expressions* such as *as far as the, greater than or equal, as well as the*. This is the least common category of noun phrases with only 4 types and 166 tokens. Some example expressions are as follows:

(56) External factors play a big role **as much as** internal factors in cheating. [TRKE2057]
(57) So, people try to make **more and more** money to live a better life [TRCU1040]
(58) There are some concepts that are **more important than** the money; these are respect, success, [TRCU1143]

The last category is *quantifier expressions* added by the researcher as a complement to Biber et. al.'s (2004) structural categorization. There are 10 types with 345 tokens in total. Some of the examples are presented below:

(59) It has **a lot of** alternatives for people from different financial [TRCU1016]
(60) They can join **all of the** elections to determine the present or [TRKE2039]
(61) today not **all of the** students but **some of them** cheat in examinations and it is [TRKE2057]

As can be seen above, the structural classifications have three main categories, each of which has 6 different sub-categories. While VP fragments have the highest number of types, dependent clause fragments have the least number of types. NP and PP expressions are the second most common fragments. In the next section, we will start presenting the analysis of the functional categories as revealed from the essays extracted from LOCNESS.

4.3.2. Structural Classification of Lexical Bundles in LOCNESS

The bundles extracted from LOCNESS and analyzed have been classified according to Biber et al.'s (2004) typology breakdown of the structures. Firstly, the overall view of the frequency counts, types and tokens of bundles found in the pertinent corpus is presented. After that, the structural classifications will be given. Figure 12 depicts the distribution of the structural types in an overall manner.

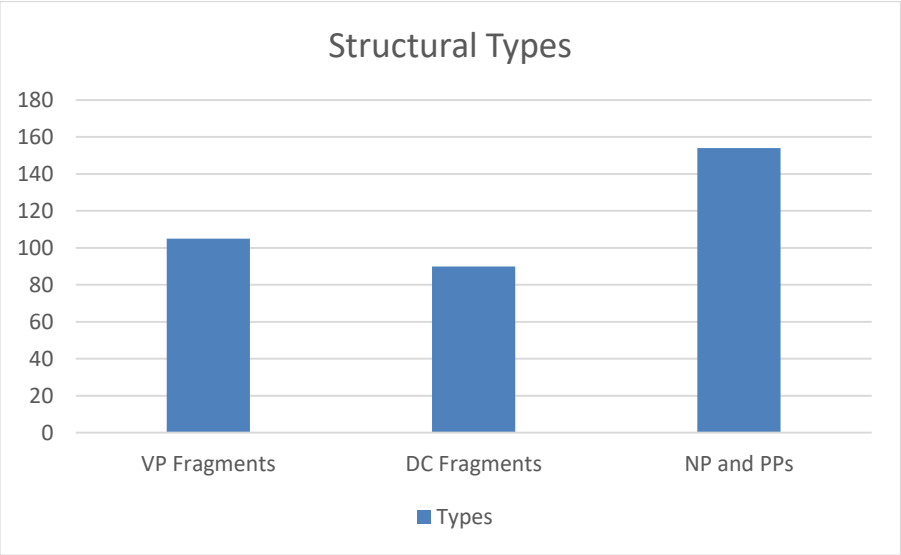


Figure 12. Distribution of structural types in LOCNESS

As can be seen from the figure, the most common structures belong to the noun phrase and prepositional phrase structures. The VP fragments follow them as the second most frequent structures. And the least frequent bundles used are seen to be the dependent clause fragments by NS in argumentative essays. When we look at the tokens of these types, the NP and PP expressions account 43% of all the tokens in the structure types. The highest number of tokens in noun phrases category belongs to the *noun phrase with of-phrase fragments.* Prepositional phrase and other noun phrase expressions follow the of-phrase fragments, respectively. The detailed summary of the structural categorizations may be seen in the following table.

Table 21.

Structural Classification of Lexical Bundles in LOCNESS

STRUCTURE	**TYPES**	**%**	**TOKENS**	**%**
1. Verb phrase fragments				
a. 1st/2nd/3rd person pronoun + VP fragment	61	17	858	18
b. Discourse marker + VP fragment	-		-	
c. Verb phrase with non-passive verb	31	12	342	11
d. Verb phrase with passive verb	12	3	142	3
e. Yes/no question fragments	-		-	
f. WH-question fragments	1	1	8	0.1
2. Dependent clause fragment				
a. 1st/2nd/3rd person pronoun + dependent clause fragment	8	2	87	2
b. WH-clause fragments	6	1	54	1
c. *If*-clause fragments	4	1	37	0.9
d. (copula be+adjective/noun phrase) *To*-clause fragments	48	10	613	8
e. *That*-clause fragments	12	3	188	4
f. Other adverbial clause fragments	12	3	153	4
3. Noun phrase and prepositional phrase				
a. Noun phrase with of-phrase fragment	56	17	822	19
b. Noun phrase with another post-modifier fragment	7	2	112	2
c. Other noun phrase expressions	35	11	550	11
d. Prepositional phrase expressions	40	12	548	12
e. Comparative expressions	5	1	52	1
f. Quantifier expressions	11	3	163	3
TOTAL	**349**	**100**	**4729**	**100**

Table 22 represents the structural categorization of LBs with their type and token frequencies. In total, there are 349 types and 4729 tokens in the whole corpus.

The first category which belongs to the verb phrase fragments, hosts 105 types and 1350 tokens constituting 33% of all the bundles in LOCNESS. VP fragments are the second most common structures used by the native speakers. As for the sub-categories, *personal pronouns + VP Fragments* category reaches the highest frequency with 61 types

and 858 tokens among all the structures. NS have not used any discourse markers such as *I mean you, you know it was and I mean I don't* and yes-no question fragment including *are you going to, do you want to, does that make sense* within the VP fragments. And wh-question fragment has only been used 8 times in the corpus.

The second category is dependent clause fragments with 90 types and 1132 tokens in total, constituting only 20% of the whole structural classifications. *To-clause fragments* are the most common structure with 48 different types and 613 tokens within this category. *That-clause fragments and other adverbial clause fragments* were both produced 12 times each becoming the second most common structures within this category. Also, the least common structure belongs to the *if-clause fragment* with only 4 types and 37 tokens.

The noun phrase and prepositional phrase structures with 154 different bundles and 2247 tokens is the most common structures in LOCNESS. These fragments constitute 46% of the whole structural classifications. *Noun phrase with of-phrase fragments* were used with 56 different types and 822 tokens, having the second highest number of tokens in the whole categorization. Second most common types used is *prepositional phrase expressions* with 40 types and 548 tokens. The least used bundle type in this group is *comparative expressions* which has 5 types 52 tokens.

4.3.2.1. Verb Phrase Fragments in LOCNESS

There exist 6 categories within verb phrase fragments. However, native speakers have used only four of these categories in their written productions. When we focus on the structural classifications in detail, NS have used *personal pronouns + VP fragments* and *VP with non-passive verb fragments* very frequently. Table below represents all types of VP fragments with their tokens in brackets below.

Table 22.

Verb Phrase Fragments in LOCNESS

a. 1st/2nd/3rd person pronoun + VP fragment	It is not (f: 39), this is not (f: 16), this is a (f: 31), it is a/an (f: 48), this is the (f: 16), there is no (f: 35), there is a (f: 40), they do not (f: 39), I don't (f: 10), there are many (f: 24), there are some (f: 12), it is the (f: 21), it is also (f: 18), it is important (f: 11), they want to (f: 20), it does not (f: 16), it has been (f: 16), they are not (f: 16), there has been (f: 15), there have been (f: 12), it should be (f: 12), changed people's lives (f: 10), the problem is (f: 10), you cannot (f: 10), it comes to (f: 9), that can be (f: 9), there will be (f: 9), they have to (f: 9), they should be (f: 9), you do not (f: 20), he does not (f: 8), I have seen (f: 8), I am not (f: 8), it is only (f: 8), people do not (f: 8), that could be (f: 8), that is the (f: 8), the child is (f: 8), they will be (f: 8), I have been (f: 7), I think it (f: 7), I would like to (f: 7), it is possible (f: 7), one can see (f: 7), there are two (f: 7), they are more (f: 7), they can be (f: 7), this argument is (f: 7), this can be (f: 7), this is a very (f: 7), we do not (f: 7), we need to (f: 7), the united states is (f: 7), crime does not pay (f: 10), it can be+V3 (f: 14), capital punishment is (f: 8), example of this is (f: 8), people's lives is (f: 7), it would be (f: 22), there would be (f: 12), the death penalty is (f: 17), that is not (f: 7)
b. Discourse marker + VP fragment	-
c. Verb phrase with non-passive verb	is not the (f: 16), is not a (f: 22), is hard to (f: 7), is not to (f: 7), is because of (f: 7), is a major (f: 7), is the root of (f: 7), is very important (f: 7), is not only (f: 11), is a very (f: 21), is one of the (f: 16), is that the (f: 16), look at the (f: 15), do not have (f: 15), has become a (f: 10), has been a (f: 10), argue that the (f: 9), should not have (f: 9), do not want (f: 8), does not mean (f: 8), does not have (f: 8), believe that they (f: 7), take care of (f: 7), should not be+adj (f: 12), would be a (f: 15), should be the (f: 7), would not be (f: 14), would not have (f: 13), will not be (f: 14), is not the only (f: 7), would have been (f: 10)
d. Verb phrase with passive verb	Be allowed to (f: 7), can be used (f: 13), be used to (f: 11), can be seen (f: 8), involved in the (f: 8), should be allowed to (f: 8), are forced to (f: 7), considered to be (f: 7), not be allowed (f: 7), should not be+V3 (f: 30), need(s) to be +V3 (f: 27), based on the (f: 9)
e. Yes/no question fragments	-
f. WH-question fragments	what is the (f: 8)

When looked at the table, *verb phrase fragments* are the second most common structure in LOCNESS, accounting for 33% of all bundles in the corpus. This category feature 105 different types, the second widest variety among all other lexical bundles structures. Tokens of this category are 1350, which makes a T/T ratio of 7.7%. Most of their structure is within *personal pronouns+VP fragments* category with 61 types. Personal pronouns+ VP fragments comprise the 17% of all the VP fragments. However, NS have not used any *discourse markers+VP fragments* such as *I mean you, you know it was and I mean I don't* and yes/no question fragments including *are you going to, do you want to, does that make sense.*

Focusing on the details, the first category, the personal pronouns+VP fragments consists of a subject pronoun followed by a verb phrase. The VP generally extends to the beginning of a following complement clause. This category holds expressions with all kinds of pronouns including, *we, they, it, you*. There is not much difference in terms of the number of the tokens of different types. Some of the examples of this type of structures may be seen below:

(62) The only problem is that **they do not** give any data that provides [USARG_0047]

(63) **it is not** fair that millions of people [USARG_0079]

(64) As of now, **there is no** evidence available concerning how effective [USARG_0022]

(65) they can to inform the public that **we need to** abolish the death penalty. [USARG_0017]

Additionally, there are also several modals used within this category such as *it should be, you cannot, there will be, I would like to, this can be, that could be.* Some of them may be seen below.

(66) not the only solution to the problem, **it should be** a definite consideration. [USARG_0025]

(67) It win also focus on the problems **that could be** avoided with the decriminalization of [USARG_0077]

(68) States at the age of eighteen, but **you can not** drink a beer like all [USARG_0163]

(69) might give them the hope that maybe **there will be** a cure before they die. [USARG_0038]

Another category which has many different types is the verb phrase with non-passive verb fragments, phrases plus a part of a following complement phrase or clause. Some of the types are *is one of the, have come to, have to do, take care of, would be able to.* Examples of this category may be seen below.

(70) craves violence and capital punishment **is one of the** only legal means of [USARG_0004]

(71) I **have come to** the conclusion that for psychological, [USARG_0125]

(72) Spencer would **argue that the** people and other welfare recipients [USARG_0079]

Also, *verb phrase with passive verb* fragments which has 12 types in total. Passive usages of "can" and "should" modals have been separated from "it can be+adj" fragments. Some passive structures found in the corpus can be seen below.

(73) Did you know that marijuana **can be used** to slow down certain eye [USARG_0142]

(74) Students **need to be** taught by professors when they [USARG_0151]

(75) This argument is mostly **based on the** legal aspects of adoption. [USARG_0044]

And the last category belongs to the wh-question fragments which have only been used 8 times with one single type *what is the* in the corpus. The examples of which are as such:

(76) surrounding this issue leaving people to wonder, **what is the** right thing to do? [USARG_0014]

(77) is known all over the world, but **what is the** real significance of his job? [USARG_0137]

(78) surrounding this issue leaving people to wonder, **what is the** right thing to do? [USARG_0014]

The next section will analyze the dependent clause fragments in detail.

4.3.2.2. Dependent Clause Fragments in LOCNESS

Native speakers have produced all of the sub-categories of dependent clause fragments. However, when it comes to the sorting process, some of the bundles could not be inserted into any of the existing categories. That is why, again, another category has been added to the existing classification as *other adverbial clause fragments*. The same category available in both TICLE and LOCNESS corpora. Table 24 indicates the categories as well as tokens of each type below.

Table 23.

Dependent Clause Fragments in LOCNESS

a. 1st/2nd/3rd person pronoun + dependent clause fragment	I feel that (f: 20), I believe that (f: 12), I think that (f: 9), people feel that (f: 8), it is obvious that (f: 9), they feel that (f: 7), those who are (f: 12), people who are (f: 10)
b. WH-clause fragments	when it comes (f: 12), what they are (f: 10), when it comes to (f: 9), when I was (f: 8), when they are (f: 8), who have been (f: 7)
c. *If*-clause fragments	if you are (f: 11), if they are (f: 10), if there is (f: 7), if a person (f: 9)
d. *To*-clause fragments (copula be+adjective/noun phrase)	to be able to (f: 50), in order to (f: 50), not be able to (f: 11), being able to (f: 7), to be a (f: 28), to be more (f: 12), to be the (f: 23), to have a (f: 22), to have an (f: 9), to go through (f: 9), to prove that (f: 9), to stay home (f: 9), to take the (f: 9), to make a (f: 17), to make the (f: 13), to say that (f: 16), to deal with (f: 15), to go to (f: 14), to do with (f: 12), not have to (f: 12), would have to (f: 12), would like to (f: 11), do not want to (f: 10), to be in (f: 8), to believe that (f: 8), to get the (f: 8), to have sex (f: 8), to keep the (f: 8), to think that (f: 8), to try to (f: 8), to live in (f: 7), to look at (f: 7), to pay for (f: 7), to support the (f: 7), should be able to (f: 7), are trying to (f: 13), seem(f: s) to be (f: 19), have come to (f: 9), is going to (f: 9), have to do (f: 8), have to go (f: 8), have the right to (f: 8), are willing to (f: 7), will be able to (f: 8), will have to (f: 8), would be able to (f: 8), have to be (f: 23), will continue to (f: 14)
e. *That*-clause fragments	that it has (f: 8), that they can (f: 8), that they would (f: 8), that do not (f: 8), that there are (f: 10), that they have(f: 10), that they were (f: 10), that they will (f: 10), that there is (f: 13), that this is (f: 13), that it is (f: 47), that they are (f: 33)
f. Other adverbial clause fragments	because it is (f: 19), because they are (f: 19), and that is (f: 9), and this is (f: 9), as it is (f: 8), that they may (f: 7), no matter what (f: 9), so that they (f: 9), whether or not (f: 19), as long as (f: 22), and it is (f: 10), but it is (f: 13)

Dependent clause fragments hold 90 types with 1132 tokens in total, becoming the least common structural category in LOCNESS. In this regard, T/T ratio of this category is 7.9%. While the most common group is to clause fragments, the least produced bundles category is *if*-clause fragments with only 4 types.

The first category is personal pronouns + dependent clause fragments which starts

with any pronouns followed by dependent phrases. This category is the third most common category of dependent clause fragments. Some examples of this category in our data include *I believe that, I feel that, people who are, those who are*…etc. Some examples of this category are as follows.

(79) In conclusion, **I feel that** the restoration of the "American [USARG_0031]
(80) of a particular racial or ethnic group, **it is obvious that** abortion is nothing like [USARG_0168]
(81) If we are willing to help **those who are** unable to help themselves, [USARG_0079]

Second category is *WH-clause fragments* which have two types; verb phrases followed by a wh-clause, or bundles that represent the beginning of a wh-clause. LOCNESS only consisted of the second type such as *when it comes, what they are, when it comes to, when I was*. The examples of this category may be seen below.

(82) twenty one are considered to be minors, **when it comes to** drinking alcoholic beverages in [USARG_0163]
(83) I remember last year **when I was-** interviewed by a representative of [USARG_0171]
(84) I believe **what they are** trying to say is, you [USARG_0072]

The third category belongs to the *if-clause fragments* which is the least frequent category of dependent clause fragments with 4 different types and 37 tokens in total. These types are *if you are, if they are, if there is, if a person.* Examples from LOCNESS are as follows:

(85) but hopefully **if you are** reading this, the point will [USARG_0161]
(86) the codes because they would feel as **if they are** betraying their friends. [USARG_0045]
(87) take a child away from the family **if there is** visible evidence of abuse or [USARG_0025]
(88) However, **if a person** begins to inject random needles [USARG_0111]

Fourth category is *(copula be+adjective/noun phrase) to-clause fragments* which have three types: verb phrases followed by a to-clause, bundles that represent the beginning of a to-clause and predicative adjective+to-clause. This category is the most frequent category of in the dependent clause fragments with 47 different types with 563 tokens in total, being the third most frequent in the whole structural classification of LOCNESS. The highest number of tokens belong to the "*to be able to*" and some other examples can be seen below:

(89) by the trial court's fundamental need **to be able to** enforce its own orders [USARG_0159]
(90) a lot of patience, time, and effort **to be a** good teacher. [USARG_0136]
(91) come to this, but ultimately, someone will **have to be** a responsible journalist and say [USARG_0171]
(92) I, for one, **do not want to** give up personal liberty [USARG_0171]

Another category is *that*-clause category which has two main types: those including main clause verb, and those comprising only the *that*-clause. Only the second main type has been found in LOCNESS. Some other types used are *that there is, that this is, that it is, that they are.* Other examples are as follows:

(93) The advocates realize that there is a numbing effect that will [USARG_0040]
(94) They attach significance to them and believe **that they are** a means of communication. [USARG_0078]
(95) Athletes are notified in advance **that they will** be tested. [USARG_0015]

The last category is the other adverbial clause fragments with 12 different types, the same as *that*-clause category. This category added to Biber et. al.'s (2004) structural categorization to fit some of bundles in TICLE has also been needed to use in LOCNESS. These fragments begin with adverbial clauses of *and, as, but, because,* and *so.* Some of the bundles of this category used in LOCNESS are *because it is, because they are, and that is, and this is.* Some of the examples are as such:

(96) has no idea who the crush is **because it is** a secret. [USARG_0150]
(97) has been so much controversy in society **whether or not** a doctor should be

allowed [USARG_0003]

(98) school who would stay after school for **as long as** a student needed her to [USARG_0136]

The next section will analyze the noun phrase and prepositional phrase fragments in detail.

4.3.2.3. Noun Phrase and Prepositional Phrase Fragments in LOCNESS

The third and the last structural category of Biber et al.'s (2004) structural classifications of lexical bundles are noun phrase and prepositional phrase fragments as mentioned above. The last category added by the researcher has also been used for LOCNESS. Table 25 represents the fragments with all the types as well as their tokens in brackets.

Table 24.
Noun Phrase and Prepositional Phrase Fragments in LOCNESS

a. Noun phrase with of-phrase fragment	the number of (f: 26), the amount of (f: 23), the case of (f: 7), in the case of (f: 15), the end of (f: 7), the end of the (f: 15), the beginning of (f: 7), aware of the (f: 12), beginning of the (f: 7), the use of (f: 45), because of the (f: 47), one of the (f: 45), one of the most (f: 18), invention of the (f: 29), the idea of (f: 28), out of the (f: 26), this type of (f: 23), a part of (f: 20), the root of (f: 12), way of life (f: 19), the effects of (f: 18), the rest of (f: 18), the lack of (f: 16), the issue of(f: 14), the question of (f: 14), the value of (f: 14), the age of (f: 13), the lives of (f: 13), the opponents of (f: 13), the people of (f: 12), the cost of (f: 12), both sides of (f: 11), a lack of (f: 11), the concept of (f: 11), the consequences of (f: 11), a matter of (f: 11), because of this (f: 9), first of all (f: 9), the life of (f: 10), the loss of (f: 10), the problem of (f: 10), the right/s of (f: 18), large number of (f: 9), one of these (f: 9), the rest of the (f: 9), the proponents of (f: 8), the types of (f: 8), a form of (f: 7), a member of (f: 7), a number of (f: 7), a sense of (f: 7), a way of (f: 7), the discovery of (f: 7), the history of (f: 7), the minds of (f: 7), the presence of (f: 7), the results of (f: 7)
b. Noun phrase with another post-modifier fragment	the fact that (f: 71), in which the (f: 9), the idea that (f: 9), the people that (f: 8), the problem that (f: 8), in a way that (f: 7), in which they (f: 7)
c. Other noun phrase expressions	the death penalty (f: 54), the right to (f: 51), the united states (f: 48), as well as (f: 40), men and women (f: 29), he or she (f: 26), his or her (f: 24), according to the (f: 20), due to the (f: 17), due to the fact (f: 11), such as the (f: 16), values and beliefs (f: 14), the only way to (f: 12), the people in the (f: 11), as a whole (f: 11), the power to (f: 9), going to be (f: 9), the opposite sex (f: 15), a good example (f: 7), an attempt to (f: 7) church and state (f: 7), rules and regulations (f: 7), the conflict theory (f: 7), the work place (f: 7), the united states government (f: 7), the American people (f: 8), the same as (f: 8), the general public (f: 9), the supreme court (f: 10), a strong argument (f: 10), the first amendment (f: 11), the best way to (f: 7), the ability to (f: 7), such as these (f: 7), on their own (f: 7)
d. Prepositional phrase expressions	in the united states (f: 53), part of the (f: 31), at the end of (f: 9), at the beginning of (f: 8), in favor of (f: 23), as a result of 15), in public schools (f: 36), in the past (f: 23), on the other hand (f: 22), in the future (f: 14), in today's society (f: 12), in the case (f: 20), in the century (f: 11), over the years (f: 11), in many cases (f: 10), in our society (f: 10), in some cases (f: 10), in terms of (f: 10), in the public (f: 10), in this case (f: 10), at the same time (f: 10), in the world (f: 19), up to the (f: 15), in front of (f: 9), in my opinion (f: 9), in the early (f: 9), around the world (f: 8), at the time (f: 8), in addition to (f: 8), in our lives (f: 8), in the article (f: 8), in the morning (f: 8), in this country (f: 8), amount of money (f: 7), an increase in (f: 7), along with the (f: 7), in high school (f: 7), in many ways (f: 7), in the classroom (f: 7), in the long run (f: 7), in the media (f: 7), with each other (f: 7)
e. Comparative expressions	the most important (f: 18), more important than (f: 7), more time to (f: 7), more and more (f: 10), a little more (f: 10)
f. Quantifier expressions	A great deal of (f: 10), all of the (f: 29), all of these (f: 13), most of the (f: 12), a lot of (f: 29), some of the 19, many of the (f: 17), many of these (f: 10), some of these (f: 9), most of the time (f: 7), the majority of the (f: 8)

Noun phrase and prepositional phrase expressions in LOCNESS have 154 different types with 2247 tokens in total, becoming the most common structures within the structural classifications in LOCNESS. While, these expressions constitute 46% of the whole structural classifications, T/T ratio is 6.8%. of Out of 6 categories, the noun phrase with of-phrase fragments are the most frequent bundles.

To begin with, the first category belongs to the noun phrase with of-phrase fragments which cover a range of meanings. A number of these bundles are used for describing physical identification of place, size and amount such as *the position of the, other parts of the, the size of the.* Some of these examples are as follows:

(99) Although **the number of** abortions performed is rising, it [USARG_0036]
(100) aspect of the welfare dilemma is **the large number of** children that are affected. [USARG_0079]
(101) in their social world homelessness would be **a part of** every day life. [USARG_0078]

Also, some bundles of the noun phrase with of-phrase fragments identify various abstract qualities such as *the nature of the, the use of the, the value of the*. LOCNESS also consisted some of these types of bundles as can be seen from the examples.

(102) drug-testing proposals aimed at stamping out **the use of** anabolic steroids by athletes. [USARG_0015]
(103) In "first-world" cultures around the globe, **the value of** a human being has become [USARG_0051]

Some other of these bundles within this category indicate simple existence or presence. Examples may be *the presence of* and *the existence of.* LOCNESS has the following:

(104) Human life is dependent on **the presence of** fresh water. [USARG_0076]

And the last type of the noun phrase with of-phrase fragments describes processes or events lasting over a period of time such as *the development of, the course of the*…etc. However, LOCNESS did not consist of this type of bundle.

Second category is *noun phrase with another post-modifier fragment* which are composed of a noun phrase followed by a post-modifier other than an *of*-phrase. These

have two major types: a noun phrase with a post-nominal clause fragment, and a noun phrase with a prepositional phrase fragment. Only the first type of these bundles has been found in LOCNESS of which examples may be seen below.

(105) They ignore **the fact that** blacks can't take part [USARG_0005]
(106) There are some cases **in which the** housekeeper is paid a flat [USARG_0143]
(107) that capital punishment does not benefit anyone **in a way that** imprisonment could not. [USARG_0004]

The third category belongs to the *other noun phrase expressions* with 35 different types and 550 tokens in total, becoming the third most common category of the NP and PP fragments. These expressions do not fit into any of the existing categories. Some of the examples can be seen as follows:

(108) capital punishment focuses on the idea of **the death penalty** acting as a deterrent. [USARG_0004]
(109) other states just as they compete for **the right to** collect and store garbage from [USARG_0056]
(110) war against all that has not benefited **the United States** at all. [USARG_0140]

Another category within the noun phrases is *prepositional phrase expressions* which is the second most frequent category. Prepositional phrases begin with a preposition and have two types which are prepositional phrases or prepositional phrase fragments with an *of* following the noun, and other prepositional phrases or prepositional phrase fragments. Some of the examples from LOCNESS are as such:

(111) **In the United States** alone, 400,000 orphans live in foster. [USARG_0025]
(112) are certain classes that you must take **in order to** advance to high school. [USARG_0162]
(113) Opponents of prayer **in public schools** adopt an attitude that carefully [USARG_0023]

As for the *comparative expressions* category, only 5 types with 52 tokens have been used by NS. These bundles are least used bundles within this categorization. The used ones are *the most important, more important than, more time to, more and more, a little more.* The example of these expressions are as follows:

(114) One of **the most important** benefits of drug legalization is [USARG_0140]
(115) made that a human's life is **more important than** an animal's life. [USARG_0020]
(116) Society should take more time to research the good things teachers [USARG_0136]
(117) Nebraska was getting a little more respect because they played in [USARG_0148]

The last category is *quantifier expressions* added by the researcher to Biber et. al.'s (2004) structural categorization. Some of these bundles in LOCNESS are *a great deal of, all of the, most of the, a lot of, some of these, the majority of the*…etc. Some of these examples are as follows.

(118) These points compose **the majority of the** supporter's arguments in [USARG_0007]
(119) This causes **a great deal of** tension in the government. [USARG_0146]
(120) could be nearly impossible to completely eliminate all of the different drug uses. [USARG_0134]

As can be seen above, the structural classifications have three main categories, each of which has 6 different sub-categories. While the most frequent type of bundles is noun phrase and prepositional phrase expressions, the least frequent bundles are dependent clause fragments. VP fragments are the second most common bundles in the whole structural classification. The next section will analyze the functional categories of LOCNESS in detail.

4.3.3. Comparison of the Structural Classifications in Two Corpora

When we compare the structural classifications in both native and non-native data, we get the following distribution of the structural classifications in figure below.

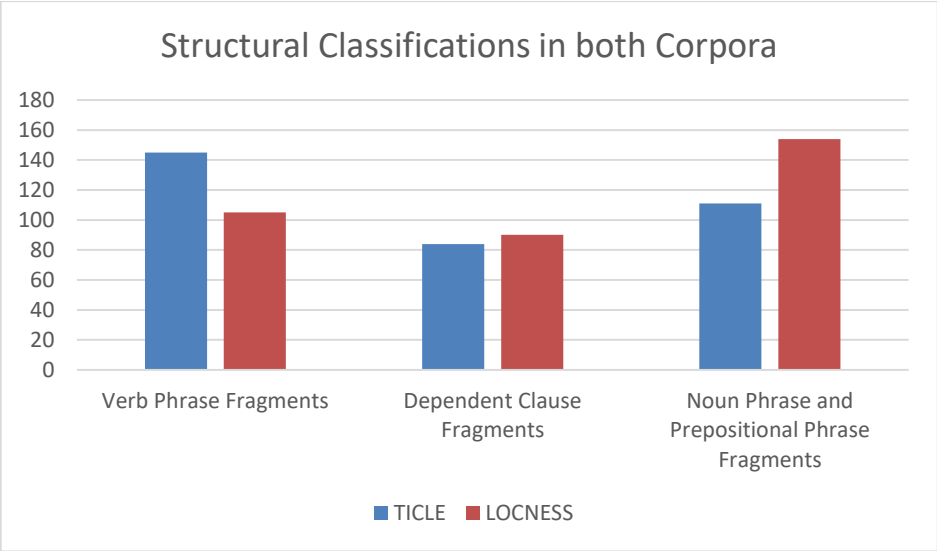


Figure 13. Structural classifications in both corpora

Looking at the structures used in both corpora reveal that while Turkish EFL learners used VP fragments as the most frequent bundles, native students used noun phrase and prepositional phrase fragments as the most frequent structures. NP and PP fragments are the second most frequent category used in TICLE, constituting 33% of the classification while VP fragments are the second most frequent structures in LOCNESS, constituting 33% of all the bundles. The dependent clause fragments are the least frequently used fragments in both corpora.

When we examine the sub-categories of these fragments in detail, personal pronouns+VP fragments is the only category constituting 42.4% of all structures in TICLE, these structures constitute only 32% of the LOCNESS. Within the VP fragments, large number of the phrases Turkish EFL learners produced report negative states which can be seen in the examples of *students/universities/people/I/you/they/we/it do(es) not* (f: 214), *they/we/people are not* (f: 90), *they/it should not be* (f: 25), *they should not* (f: 11), *I/they/you/we/he/women cannot* (f: 194)…etc. These negative structures constitute 22.6% of the personal pronouns+VP fragments category. However, this is not seen in the usages of NS. Additionally, it has been revealed in TICLE that most of the fragments within the personal pronouns+VP fragments include bundles with embedded *be*-verbs. Turkish EFL learners have also used *these* existential *there*-constructions to a great extent. The examples in our study include *there are some/many/a lot/lots of/many people/ also/so many* (f: 215), *there is a/an* (f: 108), *there is no* (f: 89), *there is no need* (f: 17), *there is not* (f: 26), *there is nothing* (f: 20). This is not the case in LOCNESS. It was revealed that NS did not produce as many existential *there*-bundles as NNS. As for the evaluative

bundles like anticipatory *it* patterns such as *it is easy to*, our study found out that Turkish L1 learners used more anticipatory *it* patterns than the natives.

As for the *"discourse marker + VP fragment"* such as *I mean you, you know it was and I mean I don't* and *"WH-question fragments"* including *what do you think, how many of you, what does that mean…etc*. have not been produced by Turkish EFL learners at all. Similarly, NS have not used *"discourse marker + VP fragment"* and *Yes/no question fragments*. They have only used one wh-question fragments in the whole data. In this regard, the usages are similar.

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

4.3.4. Functional Classification of Lexical Bundles in TICLE

Generated LBs have also been classified according to their function in the discourse using the categorization of Biber et al.'s (2004) functional classification of common lexical bundles. Three primary functions are served by lexical bundles. The first function is stance expressions which have two sub-categories: (a) epistemic stance and (b) Attitudinal/Modality stance. The second fragments of functional classifications is discourse organizers which consists of (a) topic introduction and (b) topic elaboration/clarification. And the last category is referential bundles which have (a) identification/focus, (b) imprecision, (c) specification of attributes, (d) time/place/text reference.

In the first step, the overall table including number, type and tokens of bundles are given. After that, the functional categorizations detailed are given. When looked at the distribution of these functional bundles, the functional types may be seen in Figure 14 below.

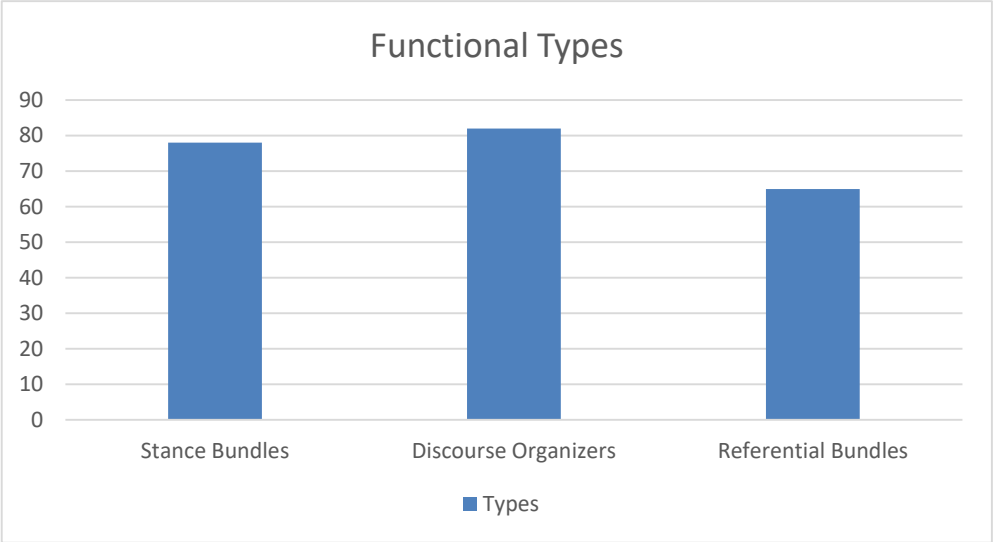


Figure 14. Distribution of functional types in TICLE

Figure demonstrates the functional types in TICLE. Even though there is not much difference between categories, the most frequent functional use by Turkish EFL learners is discourse organizers with 84 types. And then, stance bundles follow as the second frequent bundles with 78 types. And lastly, referential bundles are the least used bundles within the functional categories with 65 types. The detailed summary of the functional classifications is shown in Table 26 below.

Table 25.

Functional Classifications of LBs in TICLE

STRUCTURE	TYPES	%	TOKENS	%
1. Stance bundles				
A. Epistemic stance	14	8	334	7
B. Attitudinal/Modality stance	64	27	1434	25
2. Discourse organizers				
A. Topic introduction	33	14	641	12
B. Topic elaboration/Clarification	51	22	1541	26
3. Referential bundles				
A. Identification/ Focus	12	5	318	5
B. Imprecision	-	-	-	
C. Specification of attributes	24	11	741	13
D. Time/ Place/ Text reference	29	13	716	14
TOTAL	**227**	**100**	**5725**	**100**

As can be seen from the Table 26, the structures with their types and token are given. Out of 340 different types of lexical bundles, only 227 could be categorized according to their functions. 113 items did not have any functions, meaning that 64% of the bundles used did not serve a function in the discourse. This means that one third of the bundles used did not have any functions in the discourse. Some of the examples that did not have any functions are *the root of, the result of, is not the, the use of, aware of the, is not enough, the importance of, the right to, to have a, is not… etc.* Total number of tokens is 5725. While stance bundles and discourse organizers have two sub-categories each, referential bundles have 4 sub-categories.

The first category of functional classifications is *stance expressions* which have 78 types with 1768 tokens in total. These expressions express attitudes or assessments of certainty, framing some other proposition. They have two main types which are *attitudinal/modality stances* and *epistemic stances*. This category constitutes 35% of the whole functional classification, making it the second most common category. While *attitudinal/modality stances* are the most common functions with 64 different types and 1434 tokens while *epistemic stances* produced in TICLE have only 14 types with 334 tokens.

The second category is *discourse organizers* with 84 different types and 2182 tokens, which reflect relationship between prior and coming discourse. This category constitutes 36% of the functional categorization, having the highest percentage. Turkish EFL learners produced discourse organizer bundles more than the other kinds of functions. This category also has two kinds of bundles which are *topic introduction* and *topic elaboration/clarification*. Learners produced topic elaboration/clarification bundles more than topic introductions. There exist 51 types and 1541 tokens within the topic elaboration/clarification, becoming the second most common bundle in the functional categorization, while topic introduction has 33 types and 641 tokens in total.

The last category belongs to the referential bundles used to making direct reference to physical or abstract entities or to the texture itself. 4 different sub-categories exist within this category; identification/focus, imprecision, specification of attributes and time/place/text reference. 65 types with 1775 tokens have been used in the TICLE, constituting 29% of the functional categorization. The most used referential bundle is time/place/text reference with 29 types. However, Turkish EFL learners have not used any *imprecision bundles* including *or something like that, and stuff like that, and things like that* in the whole corpus.

4.3.4.1. Stance Bundles in TICLE

The first function served by the bundles is stance expressions which have two sub-categories: (a) epistemic stance and (b) Attitudinal/Modality stance. Sub-categories have specific functions and meanings. These types of bundles convey attitudes or assessments of certainty framing some other proposition. Stance bundles may also be classified as personal or impersonal according to the judgments and attitudes. While personal stance bundles refer to the speaker/writer with the subject 'I' as in *I don't know what*, impersonal stance bundles resemble to personal stance bundles. However, they do not refer the speaker/writer directly, such as *it is possible to* and *can be used to*.

The distribution of all types of stance bundles with their tokens in brackets have been given in Table 27 below.

,

Table 26.

Stance Bundles in TICLE

A. Epistemic stance	I believe that (f: 43), I do not (f: 39), I think the (f: 20), I think it is (f: 12), we know that (f: 16), we all know (f: 12), important thing is (f: 10), as we know (f: 13), I think that (f: 29), when we think (f: 10), if we think (f: 14), the fact that (f: 31), in my opinion (f: 70)
B. Attitudinal/Modality stance a. Desire b. Obligation/Directive c. Intention/Prediction d. Ability	I/they/you/he want(s) to (f: 139), they do not want (f: 11), when they want (f: 11), I/we/they have to (f: 72), whatever they want (f: 12), do not want (f: 18), I/we/they have to (f: 72), have to be (f: 20), we look at (f: 30), look at the (f: 33), know how to (f: 27), they need to (f: 12), important thing is (f: 10), the important thing is (f: 15), I/it/they/there will be (f: 96), they will not (f: 13), they will have (f: 10), will be a (f: 20), going to be (f: 11), to be able to (f: 26), I/they/you/we/he/women cannot (f: 194), is the most (f: 24), is the most important (f: 21), you/we can see (f: 44), we can give (f: 11), it/this/they can be (f: 73), we can say (f: 23), I/we can say that (f: 27), you/they can do (f: 33), I can say that (f: 12), it is possible (f: 12), it is impossible (f: 12), can be done (f: 17), can be used (f: 13), cannot be+V3 (f: 52), should not be (f: 33), cannot be+adj (f: 25), should not be+adj (f: 17), will be a (f: 20), should be given (f: 17), should not be (f: 33), would like to (f: 11)

The table shows the stance bundles used in TICLE in detail. Stance bundles are the second most common bundles used by Turkish EFL learners, constituting 35% of the whole functional categorization. While attitudinal /modality stance bundles have 61 types with 1362 tokens, epistemic stances feature 17 types with 406 tokens in total.

Epistemic stance bundles express the knowledge state of the information that follows. Most of them being personal, epistemic bundles may convey uncertainty such as *I don't think, I don't know what;* or a sense of imprecision like *stuff like that* and *what do you think.* As for the impersonal epistemic bundles, degrees of certainty are expressed such as *in the fact that* and *are more likely to be*. Some of the examples of these bundles from TICLE are as follows:

(121) **I believe that** euthanasia should be legalized albeit be [TRCU1176]

(122) about the capital punishment should be **the fact that** murder is a crime punishable by [TRCU1103]

(123) However, I think, the most difficult but the most humanely [TRKE2057]

Attitudinal/modality stance bundles express writers' or speakers' attitude toward the content of what follows. These bundles are mostly personal, which express speakers' attitudes towards actions. Examples of this type are *I am not going to* and *I want you.* These bundles divide into subcategories*: desire, obligation/directive, intention/prediction,* and *ability*. TICLE features 61 types with 1362 tokens of these bundles, becoming the most common bundle of all the functional categorization.

When looked at the sub-categories in detail, desire bundles frame self-motivated desires and wishes or ask about other participants' desires. Examples include *I would like to, what I want to do, I don't want to, what I want to do.* Some of the examples from TICLE may be seen below:

(124) Furthermore, **I want to** deal with two genders [TRCU1005]
(125) I **would like to** emphasize that it is really [TRCU1078]
(126) at that moment they can do **whatever they want** because of having money. [TRCU1133]

Second sub-category, *obligation and directive* bundles have a second-person pronoun. They express obligations and directions. *I want you to, you need to know, you might want to, you have to do,* and *take a look at* are some bundles in this group, some of which can be seen from TICLE below.

(127) **They need to** feel confident in themselves and their [TRKE2015]
(128) **I have to** be good at linguistics, I have to know the cultures of different countries [TRCU1129]
(129) whole day so that they could afford what **they need to** survive, and they do housework, prepare [TRCU1102]

Also, being overtly personal, *intention/prediction* bundles indicate the speaker/writer's intention to do something such as *what we're going to, is going to* (the impersonal bundle). Examples from our data may be seen below:

(130) I am **going to be** a teacher and their social statue [TRCU1136]

(131) money is their master and they are unconsciously **going to be** the slaves of it. [TRCU1041]

And lastly, *ability* bundles express the ability to do or accomplish something. Some of these bundles include *to come up with* or *to be able to*. Bundles found in our data are as such:

(132) Perhaps, the reason is that **to be able to** get the necessary things to [TRCU1056]

(133) When we look at history **it is possible** to count a few women [TRKE2004]

(134) You **can be banned** by some censorship institutions even [TRKE2024]

4.3.4.2. Discourse Organizers in TICLE

The second fragments of functional classifications are *discourse organizers*. This type of bundles organizes the flow of ideas like relationships between prior and coming discourse. They help to introduce topics, elaborate, and clarify topics. Examples include *I want to talk about, if you look at* or *going to talk about*. These types of bundles have two major functions as topic introduction/focus and topic elaboration/clarification. The distribution of all types of discourse organizers with their tokens in brackets have been given below.

Table 27.

Discourse Organizers in TICLE

A. Topic introduction	when we look at (f: 21), when we think (f: 10), when they are (f: 21), when it is (f: 17), when they want (f: 11), whatever they want (f: 12), what they have (f: 10), how to use (f: 21), what to do (f: 17), to find a (f: 22), to make the (f: 21), to be able to (f: 26), to sum up (f: 40), to be a (f: 39), to live in (f: 22), to say that (f: 20), to go to (f: 24), to have a (f: 51), if you/they have (f: 47), if we think (f: 14), if they are (f: 21), if they do (f: 17), if you are (f: 26), if it is (f: 15), if we look (f: 12), if they do not (f: 12), if they do (f: 17), if there are (f: 10), if there is no (f: 10), if there is (f: 25), if there is a (f: 13), do you think (f: 17)
B. Topic elaboration/Clarification	on the other hand (f: 119), it means that (f: 19), this means that (f: 13), that it is (f: 64), that there is/are (f: 73), that they are (f: 49), that they can (f: 27), that women are (f: 22), that they have (f: 21), to say that (f: 20), for this reason (f: 21), for example the (f: 23), as you see (f: 86), as I mentioned (f: 14), as I said (f: 13), as it is (f: 13), as we know (f: 13), in order to (f: 119), in order not to (f: 12), for this reason (f: 21), because it is (f: 22), because they are (f: 31), because of the (f: 58), because of this (f: 27), in other words (f: 21), in this way (f: 10), and they are (f: 33), and it is (f: 27), and do not (f: 31), and they can (f: 24), and so on (f: 20), because it is (f: 22), however it is (f: 11), in fact it is (f: 12), since they are (f: 13), although it is (f: 15), because they are (f: 31), so it is (f: 13), so there is (f: 13), so they can (f: 13), but it is (f: 29), but it is not (f: 15), but there is (f: 20), but they are not (f: 109), as you see (f: 86), as I mentioned (f: 14), as I said (f: 13), as it is (f: 13), as we know (f: 13), as well as (f: 20)

Discourse organizers are the most common bundles used in the TICLE, constituting 36% of the whole functional categorization. While *topic introduction* bundles have 33 types with 661 tokens, *topic elaboration/clarification* bundles have 50 types with 1521 tokens.

As for the sub-categories, *topic introduction/focus* bundles introduce a new topic. They may also express intention and desire. For example, *I want to talk about, if we look at, what I want to do* are the bundles in this group. TICLE has some of these bundles which can be seen below.

(135) **If we look at** this subject from another aspect, [TRCU1038]
(136) Secondly, **when we think** about phycological situation and future life [TRCU1158]
(137) **Do you think** female and male people have common [TRKE2028]

Topic elaboration/clarification bundles give additional explanation or clarification. Some examples include *you know I mean, was going to say.* These kinds of bundles may also be used to compare and contrast, such as *as well as the* and *on the other hand.* Examples of these types of bundles are as follows in TICLE.

(138) **On the other hand**, some governments are producing and [TRCU1048]
(139) the others and neglect social and ethic rules, **it means that** he earns money in wrong ways. [TRCU1051]
(140) This would mean **that it is** acceptable for an individual to die [TRCU1173]

4.3.4.3. Referential Bundles in TICLE

The third and the last category belongs to the referential bundles which comprise a large number and range of lexical bundles, usually referring to the textual context, physical or abstract entities as in *the nature of the* and *that's one of the.* This type has four sub-categories: identification/focus, imprecision, specification of attributes, and time/place/text reference. All of them serve different purposes. The distribution of all types of referential bundles in TICLE with their tokens in brackets have been given below.

Table 28.

Referential Bundles in TICLE

A. Identification/ Focus	the people who (f: 67), many people who (f: 16), a person who (f: 26), some people who (f: 14), the one who (f: 20), one of them (f: 17), one of them is (f: 13), is one of the (f: 40), this is a/the (f: 60), this is not (f: 29), that is the (f: 16)
B. Imprecision	-
C. Specification of attributes a. Quantity specification b. Intangible framing c. Tangible framing	as much as (f: 20), more important than (f: 15), more and more (f: 22), the most important (f: 109), a lot of (f: 113), a lot of people (f: 10), as a result (f: 47), as a result of (f: 32), in terms of (f: 18), all of them (f: 32), one of them (f: 17), most of them (f: 29), some of them (f: 35), some of the (f: 23), all of the (f: 45), most of the people (f: 21), all the people (f: 20), the rest of (f: 17), every part of (f: 21), the result of (f: 20), the root of (f: 20), according to the (f: 41), according to me (f: 14)
D. Time/ Place/ Text reference a. Place reference b. Time reference c. Text-deixis d. Multi-functional reference	in the past (f: 41), in real life (f: 16), in the exams (f: 28), in the family (f: 27), at this point (f: 21), in the future (f: 36), in our society (f: 29), in front of (f: 15), in front of the (f: 11), in the society (f: 49), all over the world (f: 32), in the world (f: 108), around the world (f: 12), all around the world (f: 18), in our country (f: 57), at the end (f: 24), at the same time (f: 24), in this century (f: 12), in the house (f: 18), in the universities (f: 15), in the real world (f: 16), in this situation (f: 16), at the beginning (f: 13), in the end (f: 13), in our lives (f: 14), in a society (f: 11), in the field (f: 12), in the life (f: 12), for a long time (f: 16)

Referential bundles are the least most common bundles used by Turkish EFL learners, constituting only 29% of the whole functional categorization. There exist 65 types with 1775 tokens in total. The second category which is *imprecision* bundles which signal imprecise reference or there may be similar references which may be provided as in *or something like that* and *things like that* have not been found in TICLE.

The first category is *identification/focus* bundles mostly characterize the subsequent proposition as important and examples are *and this is a*, *one of the things and that's one of the.* Also, identification/focus bundles mostly have discourse organizing function even though Biber, Conrad and Biber (2004) categorize them as referential expressions. Some examples are listed below.

(141) I believe that most of **the people who** apply to euthounasia can not have [TRCU1076]

(142) **This is a** remarkable aspect of nuclear energy. [TRCU1114]

(143) source of air pollution is carbon dioxide, which **is one of the** major pollutants in the atmosphere [TRCU1009]

Another subcategory of referential bundles, *specification of attributes*, identifies specific qualities of the noun that follow. While *quantity specification* ones signal quantity or amount including *have a lot of, a little bit of, a little bit about;* others *(tangible framing attributes)* describe the size and form of the head noun like *in the form of* and *the size of the.* This type has not been produced in TICLE corpus. Some others (*intangible framing attributes*) describe abstract characteristics as in the examples *in terms of, on the basis of, the nature of the*. Some of the examples of this type from TICLE may be seen below.

(144) We sent **a lot of** astronauts and spaceships to the space [TRCU1031]

(145) have to be dependent especially **in terms of** money on their family. [TRME3001]

(146) die can determine to apply to euthounesia as **the result of** a violant pain or a sudden [TRCU1076]

(147) What we mean is, now money is much **more important than** anything you can have, even more [TRCU1050]

Lastly, *time/place/text-deixis* bundles refer to particular time, place and location in the text. Examples of this type are *in the United States, the end of the year, at the same time, the beginning of the, at the end of…etc.* TICLE contains time, place and multi-functional references, however, there is not any time deixis bundles such as *shown in figure N, as shown in figure* used. Some examples from TICLE are as follows.

(148) correct when those people compare the conditions **in the past** and at present. [TRKE2003]

(149) out about every good or bad events happened **in the world** and they began to react for TRCU1033]

(150) Although there are selfish people in the world, **at the same time**, there are

still sensitive people. [TRKE2023]

4.3.5. Functional Classification of Lexical Bundles in LOCNESS

After analyzing the generated LBs according to their structures, these sequences have also been classified according to their functions in the discourse using the categorization of Biber et al.'s (2004) functional classification of common lexical bundles. The details of the functional classifications have been mentioned above (p.?). In the first step, the overall table including number, type and tokens of bundles are given. After that, the functional categorizations detailed are given. When looked at the distribution of these functional bundles, the functional types may be seen in Figure 15 below.

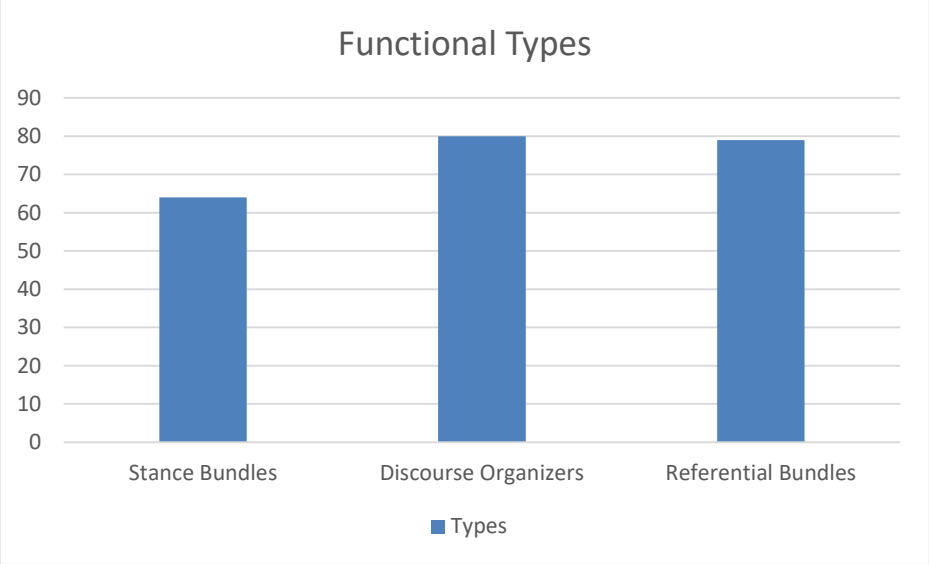


Figure 15. Distribution of functional types in LOCNESS

Figure demonstrates the functional types in LOCNESS. First, *the discourse organizers* have the highest number of types found in LOCNESS. Secondly, *referential bundles* follow in the second place. And then, *stance bundles* come having the third place among types. There is not much significance in the use of functions of the bundles NS have used. The table below gives a detailed summary of the categories.

Table 29.

Functional Classifications of LBs in LOCNESS

STRUCTURE	TYPES	%	TOKENS	%
1. Stance Bundles				
A. Epistemic stance	15	5	187	6
B. Attitudinal/Modality stance	49	23	617	21
2. Discourse Organizers				
A. Topic introduction	28	12	260	8
B. Topic elaboration/Clarification	52	23	705	24
3. Referential Bundles				
A. Identification/ Focus	16	7	330	11
B. Imprecision	-	-	-	
C. Specification of attributes	24	10	358	12
D. Time/ Place/ Text reference	39	17	479	16
TOTAL	**223**		**2936**	

As can be seen from the table, out of 349 bundles, only 223 bundles could have been categorized into the functional categories. 123 bundles did not have any functions, meaning that 64% of the bundles used did not serve a function in the discourse. Some of the examples that did not have any functions are *the general public, the power to, to take the, to stay home, a way of, an increase in, amount of money, a member of, is not to, the ability to, the work place, the history of, the minds of, the presence of, rules and regulations…etc.* The number of these categorized functional bundles' tokens is 2936. While stance bundles and discourse organizers have two sub-categories each, referential bundles have 4 sub-categories.

The first category of functional classifications is *stance expressions* which have 64 types with 804 tokens in total, making it the least most common category. They constitute only 28% of the functional categorization. As for the sub-categories, *attitudinal/modality stances* are the second most common functions with 49 different types and 573 tokens in the whole functional categorization while *epistemic stances* produced by NS have only 15 types with 208 tokens.

The second category is *discourse organizers* which reflect relationship between prior and coming discourse. LOCNESS has with 80 different types and 965 tokens, constituting 35% of the functional categorization. This category becomes the most

common category in functional classifications. This category also has two sub-categories which are *topic introduction* and *topic elaboration/clarification*. Native speakers produced topic elaboration/clarification bundles more than topic introductions. There exist 52 types and 705 tokens within the topic elaboration/clarification, becoming the most frequent bundle in the functional categorization, while topic introduction has 28 types with 260 tokens in total.

The last category belongs to the referential bundles. These expressions are used for making direct reference to physical or abstract entities or to the texture itself. 4 different sub-categories exist within this category; identification/focus, imprecision, specification of attributes and time/place/text reference. 79 types have been used in the LOCNESS, constituting 34% of the functional categorization. The most used referential bundle is time/place/text reference with 39 types. In addition, *imprecision bundles* such as *or something like that, and stuff like that, and things like that* have not been used by NS at all.

4.3.5.1. Stance bundles in LOCNESS

The first function served by the bundles is stance expressions which have two sub-categories: (a) epistemic stance and (b) Attitudinal/Modality stance. Sub-categories have specific functions and meanings. These types of bundles convey attitudes or assessments of certainty framing some other proposition.

The distribution of all types of stance bundles with their tokens in brackets have been given below.

Table 30.

Stance Bundles in LOCNESS

A. Epistemic stance	I believe that (f: 12), I think it (f: 7), I think that (f: 9), the fact that (f: 71), I feel that (f: 20), not be able to (f: 11), being able to (f: 7), is going to (f: 9), be used to (f: 11), is very important (f: 7), one can see (f: 7), to believe that (f: 8), in my opinion (f: 9)
B. Attitudinal/Modality stance a. Desire b. Obligation/Directive c. Intention/Prediction d. Ability	have to be (f: 23), involved in the (f: 8), should be allowed to (f: 8), are forced to (f: 7), not be allowed (f: 7), should not be+V3 (f: 30), should not be+adj (f: 12), need(s) to be +V3 (f: 27), we need to (f: 7), the opponents of (f: 13), there would be (f: 12), would have to (f: 12), would like to (f: 11), do not want to (f: 10), it is important (f: 11), they want to (f: 20), it does not (f: 16), it has been (f: 16), there has been (f: 15), there have been (f: 12), it should be (f: 12), would have been (f: 10), in my opinion (f: 9), it is obvious that (f: 9), there will be (f: 9), I have seen (f: 8), people feel that (f: 8), will be able to (f: 8), will have to (f: 8), would be able to (f: 8), are willing to (f: 7), considered to be (f: 7), I would like to (f: 7), be allowed to (f: 7), should be able to (f: 7), the best way to (f: 7), the results of (f: 7), this argument is (f: 7), this can be (f: 7), should be able to (f: 7), to be able to (f: 50), it is possible (f: 7), should not have (f: 9), would be a (f: 15), should be the (f: 7), would not be (f: 14), would not have (f: 13), will not be (f: 14), going to be (f: 9), can be used (f: 13), can be seen (f: 8), it can be+V3 (f: 14)

The table shows the stance bundles used in LOCNESS in detail. Stance bundles are the least most common bundles used by NS, constituting 28% of the whole functional categorization. Within the sub-categories, attitudinal /modality stance bundles are three times more frequent than epistemic stances.

The first sub-category of these expressions is epistemic stance bundles which express the knowledge state of the information that follows. The most frequent epistemic stances are *I believe that, I feel that, the fact that, can be used* in LOCNESS. Some of the concordance examples of this category may be seen below.

(151) I greatly value all of humanity and **I believe that** all people should learn to [USARG_0084]

(152) **I feel that** the Bible is very real [USARG_0006]

(153) often stand by its use based on **the fact that** it is provided for in [USARG_0012]

The second sub-category is attitudinal/modality stances that express writers' or speakers' attitude toward the content of what follows. Some of the examples found in LOCNESS are *have to be, do not want to, should not be+V3, would like to, would have to, we need to*…etc. These bundles divide into 4 subcategories*: desire, obligation/directive, intention/prediction,* and *ability*. These expressions are the most common category in LOCNESS, constituting 35% of the whole categorization.

Looking at the types of attitudinal/modality stances, the first one, desire bundles, frame self-motivated desires and wishes or ask about other participants' desires. Examples include *I would like to, what I want to do, I don't want to, what I want to do.* Some of the examples from LOCNESS are as follows:

(154) Life is considered unquestionable, but **I would like to** give it more scrutiny. [USARG_0037]

(155) I also **do not want to** suffer; a suffering that [USARG_0030]

Second type of these expressions are *obligation and directive* bundles which express obligations and directions. *I want you to, you need to know, you might want to, you have to do,* and *take a look at* are some bundles in this group, some of which have been used in LOCNESS as well.

(156) Therefore, **we need to** come to some sort of [USARG_0038]

(157) She **will have to** live knowing that she is [USARG_0132]

(158) Proper actions **need to be** taken so that the government [USARG_0152]

Another type of attitudinal/modality expressions is *intention/prediction* bundles which indicate the speaker/writer's intention to do something such as *what we're going to, is going to* (the impersonal bundle). Examples from our data may be seen below:

(159) There is always **going to be** controversy for since we live [USARG_0145]

(160) of schools and if the program is **going to be** successful, it needs support

from [USARG_008]

The last type of this category is *ability* bundles. These express the ability to do or accomplish something. Some of these bundles found in LOCNESS can be used, can be seen, to be able to…etc. some of which may be seen below.

(161) feelings that teenagers must overcome in order **to be able to** use those phrases. [USARG_0043]
(162) The radio **can be used** to communicate ideas and under [USARG_0120]
(163) While **it can be** demonstrated that both sides of [USARG_0004]

4.3.5.2. Discourse Organizers in LOCNESS

Discourse organizers are the second fragments of functional classifications which consists of (a) topic introduction and (b) topic elaboration/clarification. This type of bundles organizes the flow of ideas like relationships between prior and coming discourse. They help to introduce topics, elaborate, and clarify topics. Some examples from LOCNESS are *in order to, on the other hand, when they are, because it is*…etc.

The distribution of all types of discourse organizers with their tokens in brackets have been given in Table 32 below.

Table 31.

Discourse Organizers in LOCNESS

A. Topic introduction	when it comes (12), what they are (10), when it comes to (9), when I was (8), when they are (8), who have been (7), to look at (7), if you are (11), if they are (10), if there is (7), if a person (9), what is the (8), to take the (9), to make a (17), to make the (13), to say that (16), to deal with (15), to be in (8), to believe that (8), to get the (8), to have sex (8), to keep the (8), to think that (8), to try to (8), to live in (7), to look at (7), to pay for (7), to support the (7)
B. Topic elaboration/Clarification	in order to (50), on the other hand (22), to be a (28), to be more (12), to be the (23), to have a (22), to have an (9), to go through (9), to prove that (9), to stay home (9), to take the (9), to make a (17), to make the (13), to say that (16), to deal with (15), to go to (14), to do with (12), not have to (12), would have to (12), to be in (8), to get the (8), to have sex (8), to keep the (8), to think that (8), to try to (8), to live in (7), to pay for (7), to support the (7), because it is (19), because they are (19), and that is (9), and this is (9), as it is (8), that they may (7), no matter what (9), so that they (9), whether or not (19), as long as (22), and it is (10), but it is (13), that it has (8), that they can (8), that they would (8), that do not (8), that there are (10), that they have(10), that they were (10), that they will (10), that there is (13), that this is (13), that it is (47), that they are (33)

Discourse organizers are the most common bundles used in the LOCNESS with 80 different types and 965 tokens, constituting 35% of the functional categorization. While *topic introduction* bundles have 28 types, *topic elaboration/clarification* bundles have 52 types.

The first sub-category is *topic introduction/focus* bundles which introduce a new topic, express intention, and desire. For example, *I want to talk about, if we look at, what I want to do* are the bundles in this group. LOCNESS has some of these bundles which can be seen below.

(164) **If a person** chooses to do whatever 'in [USARG_0165]

(165) twenty one are considered to be minors, **when it comes** to drinking alcoholic beverages in [USARG_0163]

(166) How are you supposed to have fun **if you are** always asked to tone it

[USARG_0131]

The second sub-category belongs to the *topic elaboration/clarification bundles* which give additional explanation or clarification, comparing as well as contrasting. Some examples of these types of bundles are from LOCNESS are *on the other hand, in order to, that it is, as long as*…etc. Some other example concordance lines are below.

(167) **On the other hand,** when something is unethical**,** [USARG_0020]

(168) it doesn't matter about relationships **as long as** they get their money and [USARG_0026]

(169) but it is quite obvious **that it is** a very significant problem. [USARG_0009]

4.3.5.3. Referential Bundles in LOCNESS

The third and the last category belongs to the referential bundles which comprise a large number and range of lexical bundles. These expressions constitute 34% of the functional categories with 79 types. They are the second most frequent expressions in LOCNESS. The distribution of all types of referential bundles in LOCNESS with their tokens in brackets have been given below.

Table 32.

Referential Bundles in LOCNESS

A. Identification/ Focus	those who are (f: 12), people who are (f: 10), is one of the (f: 16), this is not (f: 16), this is a (f: 38), it is a/an (f: 48), this is the (f: 16), there is no (f: 35), there is a (f: 33), one of the (f: 45), one of the most (f: 18), one of these (f: 9), the rest of the (f: 9), such as these (f: 7), the people in the (f: 11), in which they (f: 7)
B. Imprecision	-
C. Specification of attributes a. Quantity specification b. Intangible framing c. Tangible framing	the most important (f: 18), more important than (f: 7), more time to (f: 7), more and more (f: 10), a little more (f: 10), great deal of (f: 10), all of the (f: 29), all of these (f: 13), most of the (f: 12), many of the (f: 17), many of these (f: 10), some of these (f: 9), most of the time (f: 7), as well as (f: 40), as a result of (f: 15), a lot of (f: 29), some of the (f: 19), in terms of (f: 10), the majority of the (f: 8), as a whole (f: 11), large number of (f: 9), the rest of the (f: 9), the number of (f: 26), the amount of (f: 23)
D. Time/ Place/ Text reference a. Place reference b. Time reference c. Text-deixis d. Multi-functional reference	in the united states (f: 53), part of the (f: 31), at the end of (f: 9), at the beginning of (f: 8), in favor of (f: 23), in public schools (f: 36), in the past (f: 23), in the future (f: 14), in today's society (f: 12), in a way (f: 11), in the case (f: 20), in the century (f: 11), over the years (f: 11), in many cases (f: 10), in our society (f: 10), in some cases (f: 10), in the public (f: 10), in this case (f: 10), at the same time (f: 10), in the world (f: 19), up to the (f: 15), in front of (f: 9), in the early (f: 9), around the world (f: 8), at the time (f: 8), in addition to (f: 8), in our lives (f: 8), in the article (f: 8), in the morning (f: 8), in this country (f: 8), along with the (f: 7), in high school (f: 7), in many ways (f: 7), in the classroom (f: 7), in the long run (f: 7), in the media (f: 7), with each other (f: 7)

As can be seen from the table, these types of expressions have four sub-categories: identification/focus, imprecision, specification of attributes, and time/place/text reference. All of them serve different purposes.

To start with, the first sub-category is *identification/focus* bundles. Some examples are *and this is a, one of the things and that's one of the.* Also, identification/focus bundles mostly have discourse organizing function even though Biber, Conrad and Biber (2004) categorize them as referential expressions. Some

examples from LOCNESS are listed below.

(170) Computers are indeed **one of the** greatest inventions of the 20th [USARG_0087]
(171) **It is a** known fact that for most [USARG_0004]
(172) **There is a** concentration of African-Americans in [USARG_0078]

Specification of attributes is another subcategory of referential bundles, which identifies specific qualities of the noun that follow. There are 3 types of these expressions. The first one is *quantity specification* which signal quantity or amount as in the examples *have a lot of, a little bit of, a little bit about.* The second type is *tangible framing attributes* that describe the size and form of the head noun such as *in the form of* and *the size of the.* This type has not been produced in LOCNESS corpus. And the last type is *intangible framing attributes*. These describe abstract characteristics as in the examples *in terms of, on the basis of, the nature of the*. Some of the concordance lines from LOCNESS are given below.

(173) This would save the government **a lot of** money and make people support [USARG_0152]
(174) **The number of** Episcopalians has been in decline [USARG_0018]
(175) possibly in danger in the Simpson house **as a result of** police inaction. [USARG_0046]
(176) They claim that **the amount of** information found by testing animals [USARG_0020]

The last sub-category of referential bundles is *time/place/text-deixis* bundles which refer to particular time, place, and location in the text. Some examples of this type in LOCNESS are *in the united states, at the beginning of, in our society, in this country*…etc. Although time, place and multi-functional references have been used in LOCNESS, there is not any time deixis bundles such as *shown in figure N, as shown in figure* used. Some examples may be seen below.

(177) schools may continue to gain more popularity **in the United States**. [USARG_0023]

(178) she was living in a homeless shelter **at the time** [USARG_0044]

(179) playoff system and keep their bowl games **at the same time**. [USARG_0148]

(180) They do this because **at the beginning of** the next fiscal year [USARG_0146]

Lastly, the second category which is *imprecision* bundles which signal imprecise reference or there may be similar references which may be provided as in *or something like that* and *things like that* have not been produced by NS.

4.3.6. Comparison of Functional Classifications in Two Corpora

When we compare the productions in terms of having functions in discourse, there is not much difference between the two corpora in terms of structures having functional properties. Approximately one third of the bundles used did not have any functions in the discourse in both corpora. The classifications of both corpora may be seen in the figure below.

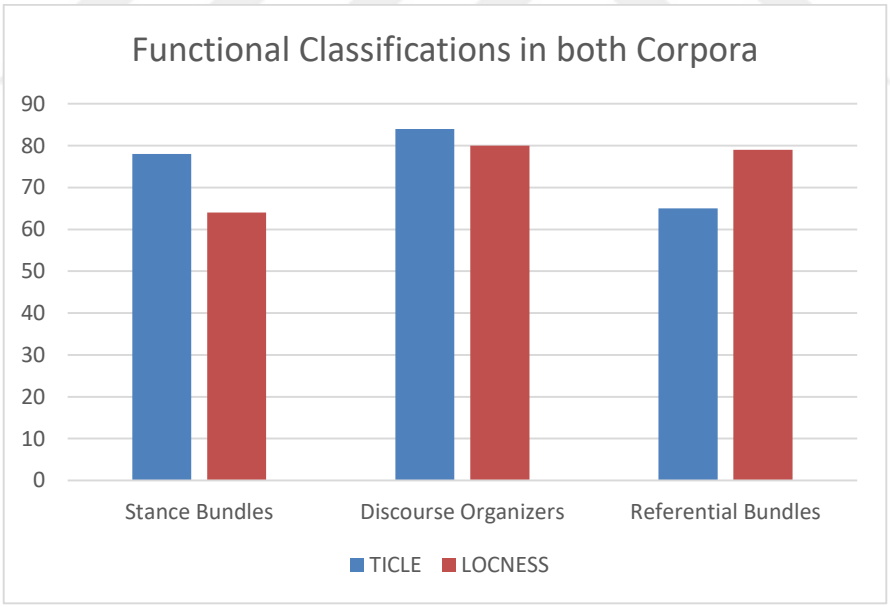


Figure 16. Functional classifications in both corpora

Turkish EFL learners (36%) and native speakers (35%) have used discourse organizers most. While stance bundles have been used as the second most frequent function in TICLE (35%), referential bundles are the second most common function in LOCNESS (34%). And the last category used by Turkish EFL learners is referential

bundles (29%), whereas the stance bundles are the least used function by the NS (28%). As for the referential bundles, these are the least frequently used expressions by Turkish EFL learners while NS used them as the second most frequent bundles. Within the referential bundles, one of the sub-categories is imprecision which include bundles like *or something like that* have not been produced in both TICLE and LOCNESS.

4.4. Findings for the Research Question 3

RQ3. How similar or different are the frequently used 3- to 4- lexical sequencies used by Turkish EFL learners from those found in native English speakers' (LOCNESS) corpora in terms of:

a. structure,
b. function, and
c. diversity in their use?

For this question, Log-likelihood analysis has been carried out in order to reveal the similarities and differences between two corpora. The results have been given and explained in detail in the following sections.

4.4.1. Log-likelihood Analysis Results

As can be seen in the results of the structural and functional categories, the frequency of the bundles is various; therefore, in order to determine whether there is an overuse or underuse, Log-likelihood analysis has been carried out. The results may be seen in the following sections.

4.4.1.1. Comparison of Structural Classifications of LBs in both Corpora

Initial LL analysis measured higher or lower frequencies of type frequencies between TICLE and LOCNESS corpora to test the overall frequency difference occurred between them. Latter LL analysis measured the token frequencies of bundles in both corpora. Table 34 below shows the LL values for the total number of types in structural classification in both corpora.

Table 33.

Log-Likelihood Results of Structural Types of LBs in both Corpora

Structure	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
1. Verb phrase fragments	145	105	0.07	0.787 +
2. Dependent clause fragments	84	90	5.54	0.019* -
3. Noun phrase and prepositional phrase fragments	111	154	24.88	0.000***-
TOTAL	**340**	**349**	**16.97**	0.000***-

As can be seen, all the structural categories have been underused in TICLE except for the VP fragments of which LL value is not significant. In total, Turkish EFL learners have underused all of the categories of the structural classification with 16.97- LL value which is quite significant. The most significant difference belongs to the noun phrase expressions with a 24.26- LL value. And then, second significant difference is the usage of dependent clause fragments with 5.84- LL ratio. The difference in the VP fragments is not significant. Underusing number of types indicates that lexical variety as well as richness is too less than NS. In order to see the total number of tokens in both corpora, we get the following results.

Table 34.

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

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

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

The table shows the number of tokens in in both corpora along with the LL differences and significance. The most significant overuse belongs to the VP fragments.

The LL value of this category is 452.05+ which is by far the highest value among all the other categories. And then, dependent clause fragments follow with 21.24+ LL value which is also overused in TICLE compared to the LOCNESS. The last significant difference is noun phrase and prepositional phrase category. Even though number of different variety of bundles used by Turkish EFL learners are far less than native students, the number of tokens is overused in all categories, meaning that the few numbers of bundles have been used numerous times by Turkish EFL students repetitively.

Table 36 shows the LL ratio of verb phrase fragment types in TICLE and LOCNESS.

Table 35.

Log-Likelihood Results of VP Fragment Types in both Corpora

Verb Phrase Fragments Structure	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
a. 1st/2nd/3rd person pronoun + VP fragment	115	61	4.92	0.027* +
b. Discourse marker + VP fragment	0	0	-	-
c. Verb phrase with non-passive verb	19	31	7.40	0.007 **-
d. Verb phrase with passive verb	10	12	1.21	0.271 -
e. Yes/no question fragments	1	0	-	-
f. WH-question fragments	0	1	-	-
TOTAL	**145**	**105**	**0.07**	**0.787 +**

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

According to the table, LL ratio measurement indicates an overuse in TICLE with an 0.07+ LL value. This means that even though there is not a significant difference between two corpora in terms of types of frequency ($p < 0.05$) in total, *personal pronouns+VP fragment and Verb phrase with non-passive verb* categories have significant LL values. *Personal pronouns+VP fragments* have 4.92+ LL value, meaning that Turkish EFL learners have overused these fragments. Some of the most frequent types of this structure in TICLE are *it is not* (f: 120), *they do not* (f: 119), *there is no* (f: 106), *there is a* (f: 86), *it is a* (f: 76), *there are some* (f: 72), *it is the* (f: 69), *they are not* (f: 65), *I want to* (f: 59), *they cannot* (f: 53), *I believe that* (f: 43). However, most frequent

examples of this sub-category in LOCNESS include *it is not* (f: 39), *this is not* (f: 16), *this is a* (f: 31), *it is a/an* (f: 48), *there is no* (f: 35), *there is a* (f: 40), *they do not* (f: 39). Even though the phrases used are similar, the frequencies are far more in TICLE than in LOCNESS as can be seen in brackets. Another excessive usage in this sub-category is the modals used by Turkish EFL learners such as; *they should not, I can say that, you can see, it can be*…etc. Total number of types used with modals in this category is 26. NS produced 11 types of modals of which examples include *it should be, you cannot, that can be, they should be, that could be, one can see, they can be, this can be, we need to, it would be…etc.* Many examples of these structures may be seen in the previous chapter in detail.

Verb phrase with non-passive verb is the other significant difference between two corpora. Turkish EFL learners have underused these types of fragments compared to the NS. Also, verb phrase with passive verb fragments have been underused in TICLE. LL calculations could not be made in yes/no and wh-question fragments since there was only one type in each in one corpus at a time. As for the discourse markers, there was not any usage in both corpora.

Table 37 shows the LL values of tokens of these fragments in TICLE and LOCNESS.

Table 36.

Log-Likelihood Results of VP Fragment Tokens in both Corpora

Verb Phrase Fragments Structure	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
a. 1st/2nd/3rd person pronoun + VP fragment	2672	858	528.55	0.000***+
b. Discourse marker + VP fragment	-	-		-
c. Verb phrase with non-passive verb	588	342	14.20	0.000***+
d. Verb phrase with passive verb	194	142	0.05	0.829 +
e. Yes/no question fragments	12	-	8.07	**0.005**+**
f. WH-question fragments	-	8	8.40	**0.004**-**
TOTAL	**3466**	**1350**	**452.05**	**0.000***+**

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

According to the LL ratio measurement, the most significant difference seems to belong to *personal pronouns+VP fragments* with 528+ LL value. Turkish EFL learners have produced these structures 3 times more frequent than the native speakers. The most frequent number of tokens belong to the first sub-category. The highest number of tokens belong to the "third person singular pronoun+auxiliary verb+noun/adj/adverb" structure. Some examples from the TICLE may be seen below:

(1) **So, it is a** communication device for people. [TRCU1032]
(2) As you see, **it is impossible** to live without money. [TRCU1056]
(3) not know the content of it or why **it is so** important to me to be a [TRCU1066]
(4) Nowadays, **it is very** popular to find friends by the [TRCU1015]
(4) **It is obvious that** their characters begin to resemble [TRKE2046]

Second significant difference belongs to the *verb phrase with non-passive verb* fragments with a 14.20+ LL value. Even though there is an underuse in terms of types of *verb phrase with passive verb fragments* compared to the NS, tokens are more frequent which indicates and overuse in TICLE. Also, fragments have been underused in TICLE. As for the LL calculations in yes/no and wh-question fragments, Turkish EFL learners have used 12 times yes/no question fragments and not used any wh-question fragments. NS have produced wh-question fragments 8 times while have not produced any yes/no question fragments. There is not any usage in the discourse markers in both corpora. Verb phrase fragments have been overused in TICLE with a 452.05+ LL value when looked at the total number of tokens.

LL Ratio of dependent clause fragment types in TICLE and LOCNESS have also been compared which can be seen below in Table 38.

Table 37.

Log-Likelihood Results of DC Fragment Types in both Corpora

Dependent Clause Fragments Structure	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
a. 1st/2nd/3rd person pronoun + dependent clause fragment	15	8	0.62	0.430 +
b. WH-clause fragments	9	6	0.05	0.823 +
c. If-clause fragments	14	4	3.37	0.066 +
d. To-clause fragments (copula be+adjective/noun phrase)	13	48	32.71	0.000 ***-
e. That-clause fragments	10	12	1.21	0.271 -
f. Other adverbial clause fragments	23	12	1.07	0.301 +
TOTAL	**84**	**90**	**5.54**	**0.019*-**

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

When looked at the table, the 4 of the sub-categories have been overused in TICLE. These are *personal pronouns+depen dent clause fragments, wh-clause fragments, if-clause fragments* and *other adverbial clause fragments.* However, *to-clause fragments* and *that-clause fragments* have been underused significantly that affect the total LL ratio. The most significant difference in this category belongs to *to-clause fragments* with 35.64- LL value, meaning that Turkish EFL learners underused these items too much compared to the NS. There are 7 common types in both corpora. These are *to be able to (50), to be a (28), to have a, to make the (13), to say that, would like to, have to do (8). Examples from both corpora can be seen below.*

(181) **it is very important to be able to use a computer to find**. [TRCU1022]

(4) educated the uneducated which will enable them **to be able to** provide a better life [USARG_0072]

(182) parents are looking forward **to have a** boy by saying that [TRCU1003]

(183) a walking womb that will help them **to have a** child. [USARG_0010]

(184) First of all, I **would like to** talk about how the education should [TRME3017]

(185) stated facts as well as opinions, I **would like to** express my opinion on the [USARG_0132]

Some of the other to-clause types in LOCNESS include *to deal with* (f: 15), *to go to* (f: 14), *to do with* (f: 12), *not have to* (f: 12), *would have to (f: 12), would like to* (f: 11), *do not want to (f: 10), to be in (f: 8), to believe that (f: 8), to get the (f: 8), to have sex (f: 8), to keep the* (f: 8), *to think that* (f: 8), *to try to* (f: 8), *to live in* (f: 7), *to look at* (f: 7), *to pay for* (f: 7), *to support the* (f: 7).

The other underused sub-category is *that-clause fragments.* LL ratio measurement indicates an underuse in dependent clause fragment types in TICLE with a 5.84- LL value which is significant.

Table 39 shows the LL values of tokens of these fragments in TICLE and LOCNESS.

Table 38.

Log-Likelihood Results of DC Fragment Tokens in both Corpora

Dependent Clause Fragments Structure	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
a. 1st/2nd/3rd person pronoun + dependent clause fragment	319	87	82.48	0.000***+
b. WH-clause fragments	140	54	18.71	0.000***+
c. If-clause fragments	239	37	112.61	0.000***+
d. To-clause fragments (copula be+adjective/noun phrase)	407	613	122.36	0.000 ***-
e. That-clause fragments	280	188	1.37	0.241 +
f. Other adverbial clause fragments	470	153	90.64	0.000***+
TOTAL	**1855**	**1132**	**30.20**	**0.000***+**

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

The token LL ratio measurements seem to be consistent with the LL ratios of types. The most significant difference is *to-clause fragments* with 200.25- LL value. This means that these structures have been extremely overused in TICLE. Another extreme overused structure is if-clause fragments compared to the NS. The other overused structures are *personal pronouns+dependent clause fragments* and *wh-clause fragments.* In total, dependent clause fragments have been overused with a 21.24+ LL ratio by Turkish EFL learners which is quite significant.

Next comparison is log-likelihood result of noun phrase and prepositional phrase fragment types in TICLE and LOCNESS.

Table 39.

Log-Likelihood Results of NP Fragment Types in both Corpora

Noun Phrase and Prepositional Phrase Fragments	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
a. Noun phrase with of-phrase fragment	14	56	40.54	0.000 ***-
b. Noun phrase with another post-modifier fragment	5	7	1.16	0.282 -
c. Other noun phrase expressions	22	35	7.92	0.005 **-
d. Prepositional phrase expressions	56	40	0.05	0.815+
e. Comparative expressions	4	5	0.59	0.444 -
f. Quantifier expressions	10	11	0.77	0.380 -
TOTAL	**111**	**154**	**24.88**	**0.000*****-

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

Table 40 indicates that Turkish EFL learners have underused 5 of the sub-categories in noun phrase and prepositional phrase structures. Only prepositional phrase expressions seem to be overused in TICLE which is not significant. The most significant belongs to the noun phrase with of-phrase fragments with 40.54- LL ratio. Some of these phrases used by native speakers are *the number of, the amount of, the use of, because of the, one of the, invention of the, the idea of, out of the, this type of.*

Second most significant difference is other noun phrase expressions with a 7.92-LL value. Common types in both corpora in this sub-category are *as well as, men and women, his or her, according to the.* Extracted examples can be seen below.

(186) All people, **men and women,** are given the opportunity of [TRCU1102]

(187) They also need to learn that **men and women** are not limited by their [USARG_0128]

(188) **According to the** article on the internet; in [TRCU1166]

(189) **According to the** author of the December 7, 1994 issue [USARG_0019]

Some of the types produced by NS are *an attempt to (7) church and state (7), rules*

and regulations (7), the conflict theory (7), the work place (7), the united states government (7), the American people (8), the same as (8), the general public (9), the supreme court (10), a strong argument (10), the first amendment (11), the best way to (7), the ability to (7), such as these (7), on their own (7). Examples from these two sub-categories may be seen in the previous chapter.

When looked at the total number of types, LL ratio measurement indicates an underuse in noun phrase and prepositional phrase fragments in TICLE with a 24.26- LL value which is extremely significant. In other words, lexical variety and richness in TICLE are less than that of NS in terms of using bundles.

Next table shows the LL values of tokens of noun phrase and prepositional phrase fragments in TICLE and LOCNESS.

Table 40.

Log-Likelihood Results of NP Fragment Tokens in both Corpora

Noun Phrase and Prepositional Phrase Structures	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
a. Noun phrase with of-phrase fragment	394	822	302.42	0.000***-
b. Noun phrase with another post-modifier fragment	143	112	0.12	0.729 -
c. Other noun phrase expressions	763	550	0.49	0.483 +
d. Prepositional phrase expressions	1363	548	163.64	0.000***+
e. Comparative expressions	166	52	34.34	0.000***+
f. Quantifier expressions	345	163	24.74	0.000***+
TOTAL	**3174**	**2247**	**4.32**	**0.038 *+**

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

As can be seen from the table, the token LL ratio measurements indicate an overuse in four of sub-categories. Even though there is an underuse in the LL ratio of types, there seems to be an overuse in the number of the tokens. Noun phrase with of-phrase fragments have the most significant difference when compared to the NS. These structures have been underused in TICLE with a 302.42- LL ratio. Another significant difference belongs to the prepositional phrase expressions which have been overused by Turkish EFL learners with 181.52+ LL value. The other overused expressions are

comparative and quantifier expressions, respectively. In total, LL ratio measurement indicates that Turkish EFL learners have significantly overused noun phrase and prepositional phrase structures with a 7.79+ LL value.

Overall, native, and non-native students produce similar LBs in their argumentative essays. Even though our study indicates some common uses in terms of LBs, we also showed that some specific bundles were favored by either the native or the non-native students, including heavy use of VP-based bundles, overuse of *existential*-there constructions, underuse of passives, significant overuse of same bundles, overgeneralizing some bundles. These features have been described in previous studies as features of learner' lexical bundle uses. To sum up, the log-likelihood test between TICLE and LOCNESS showed significant differences in both groups' use of bundles in terms of structures of the bundles. The differences of types of these bundles reveal that lexical variety and richness in TICLE are less than that of NS. When looked at the tokens of the very same bundles, there has been an extreme overuse by Turkish EFL learners, meaning that these learners have produced same bundles repeatedly.

4.4.1.2. Comparison of Functional Classifications of LBs in both Corpora

Log-likelihood analysis has been carried out in functional classifications of two corpora in order to reveal the differences between two corpora based on the normalized frequencies.

The LL ratio of functional types in TICLE and LOCNESS may be seen in Table 42 below.

Table 41.

Log-Likelihood Results of Functional Types in both Corpora

Functional Structures	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
1. Stance bundles				
a. Epistemic stance	14	15	0.92	0.337 -
b. Attitudinal/Modality stance	64	49	0.01	0.911 -
2. Discourse organizers				
a. Topic introduction	33	28	0.23	0.630 -
b. Topic elaboration/Clarification	51	52	2.43	0.119 -
3. Referential bundles				
a. Identification/ Focus	12	16	2.31	0.129 -
b. Imprecision	-	-	-	-
c. Specification of attributes	24	24	0.99	0.319 -
d. Time/ Place/ Text reference	29	39	5.76	0.016 -
TOTAL	**227**	**223**	**8.19**	**0.004 ** -**

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

According to the functional classification, all of the categories have been underused by Turkish EFL learners in terms of the LL measurement of types. This means that most of the structures Turkish EFL learners used do not have functional properties. When looked at the table, 2 of these structures have been significantly underused. The most significant difference belongs to the time/place/text reference with a 5.76- LL value, meaning that these expressions have been underused extremely compared to the native speakers. Second most significant difference is topic elaboration/clarification with a value of 2.43- which have been underused in TICLE. The other differences may not be considered significant according to LL measurement ($p < 0.05$).

As can be seen from the table, there is not much difference between the two corpora in terms of structures having functional properties. However, LL ratio measurement indicates that all three functional categories have been underused by Turkish EFL learners. When looked at the total, the difference is 8.19- LL value which is extremely significant. In other words, lexical variety, and richness in TICLE are less than that of NS in terms of using functional bundles.

The LL ratio of tokens of these functional categories may be seen in Table 43 below.

Table 42.

Log-Likelihood Results of Functional Tokens in both Corpora

Functional Structures	**Freq. in Corpus TICLE**	**Freq. in Corpus LOCNESS**	**Log-likelihood**	**Sig.**
1. Stance bundles				
a. Epistemic stance	334	187	10.46	0.001 **+
b. Attitudinal/Modality stance	1434	617	141.64	0.000 ***+
2. Discourse organizers				
a. Topic introduction	641	260	75.14	0.000 ***+
b. Topic elaboration/Clarification	1541	705	124.34	0.000 ***+
3. Referential bundles				
a. Identification/ Focus	318	330	17.09	0.000 ***-
b. Imprecision	-	-	-	-
c. Specification of attributes	741	358	48.68	0.000 ***+
d. Time/ Place/ Text reference	716	479	3.74	0.053 +
TOTAL	**5725**	**2936**	**289.77**	**0.000 ***+**

+ indicates overuse in TICLE relative to LOCNESS,
- indicates underuse in TICLE relative to LOCNESS

When looked at the LL ratio of token in the functional classification, almost all the categories have been overused by Turkish EFL learners except for the identification/focus bundles. This indicates that even though all types of these categories have been significantly underused in TICLE, the number of tokens indicates an extreme overuse. In other words, the same structures have been produced numerous times in TICLE. Both stance bundles have been overused in TICLE. The second significant difference belongs to the Attitudinal/Modality stances with 113.74+ LL ratio in the whole categorization. Turkish EFL learners have overused these stances when compared to the NS.

In the second category, the discourse organizers, the most significant difference is topic elaboration/Clarification category with a LL value of 119.42+ which means that these structures have been extremely overused in TICLE. It is also the most significant

difference in the whole functional classification. And topic introduction bundles have also been overused by Turkish EFL learners as the third most significant difference with 81.64+ LL value.

As for the referential bundles, *specification of attributes bundles* has been overused in TICLE with a 80.36+ LL value, becoming the biggest difference within this category. Identification/ Focus bundles have also been overused by Turkish learners, following as the second biggest difference in this category. The least difference belongs to the time/ place/ text reference bundles with a 5.86+ LL ratio. The only underused bundles in TICLE are identification/focus bundles in the whole functional classification.

When looked at the total, LL ratio measurement indicates an overuse in TICLE with a 319.69+ LL value. This means that there is a significant difference between two corpora in terms of the frequency of functional categories of the bundles. The overuse in TICLE relative to LOCNESS has been approved by LL calculation.

Overall, in the functional categorization of bundles, the log-likelihood test between TICLE and LOCNESS has revealed that the variety of functional types have been significantly underused by Turkish EFL learners. However, when looked at the frequency of these types, namely tokens, we can clearly see that all the categories have been overused.

4.5. Findings for the Research Question 4

RQ4. What are the semantic domains of multi-word expressions in both TICLE and LOCNESS?

In order to answer our last question, UCREL Semantic Analysis System (USAS) semantic tag sets have been used utilizing Wmatrix (Version 4) semantic analysis platform (Rayson, 2008) to identify the most salient features of the multiword choices of learners. After viewing the word frequency profile, multiword expressions have been chosen to get the MWIs in two corpora separately. When sorted by their frequency, following lists have been obtained on Wmatrix online platform.

Table 43.

The Most Frequent Multiword Expressions in TICLE on Wmatrix

Word	**Semtag**	**Freq.**	**Relative Freq.**
because_of	A2.2	232	0.12
for_example	Z5	200	0.11
have_to	S6+	177	0.09
each_other	Z8	151	0.08
according_to	Z5	143	0.08
a_lot	N5+	135	0.07
in_order_to	Z5	119	0.06
on_the_other_hand	Z4	105	0.06
university_degrees	P1	88	0.05
real_life	A5.4+	85	0.04
in_fact	A5.2+	83	0.04
in_my_opinion	Z4	70	0.04
first_of_all	N4	65	0.03
look_at	X3.4	64	0.03
i_think	Z4	63	0.03
such_a	A13.3	62	0.03
their_own	Z8	62	0.03
animal_testing	X2.4	58	0.03
so_many	N5	57	0.03
day_by_day	A6.2+	50	0.03
as_well	N5++	44	0.02
sum_up	Q2.2	41	0.02
in_addition_to	N5++	41	0.02
in_the_pas	T1.1.1	36	0.02
point_of_view	X4.1	36	0.02
all_over_the_world	M7	32	0.02
had_to	S6+	31	0.02
for_real	A5.4+	30	0.02
as_it_is	Z4	27	0.01
in_front_of	N4	26	0.01
at_the_same_time	T1.1.2	24	0.01
more_and_more	A13.3	22	0.01
at_this_point	T1.1.2	21	0.01
get_rid of	A9-	21	0.01
long_time	T1.3+	21	0.01
in_other_words	Z4	21	0.01
in_spite_of	S8-	20	0.01
and_so_on	Z4	20	0.01
in_terms_of	Z5	18	0.01
you_see	Z4	17	0.01
do_you_think	Z4	17	0.01
in_the_end	N4	13	0.01
as_i_said	Z4	9	0.00

*** Semtag representing the semantic field numbers**

After obtaining the lists, the expressions have been compared with the ones which were extracted and analyzed for the previous research questions using AntConc corpus analysis tool in order to eliminate the ones we did not have in our initial extracted list.

As can be seen from the figure, the frequency cut-off has been 6 in TICLE since there were not any matching expressions in the rest of the list with our present bundles list. There are 45 multiword expressions found in TICLE and their related semantic tag have been given. The most frequent expressions are *because of, have to, each other, for example, according to, a lot.* Sample concordance lines from the most frequent item *because of* may be seen below.

at who has right to kill someone	because of	his crime . In my opinion , nobod
body has some kind of problems ,	because of	this problems the people commit a
erefore nobody can judge someone	because of	its crime by themselves by hangin
people tried to kill themselves	because of	their love and admiration for him
and are influencing each other .	Because of	this , the two sexes must have th
of concerning about the result ,	because of	ignorance . I can say more intere
f they should be ashamed of it .	Because of	that children have been brought u
her . The students need to cheat	because of	lots of reasons . They cheat if t
e lots of murder events everyday	because of	unimportant reasons . Of course t
hose are mental disorders rising	because of	depression , stress , financial p
fore . It is a universal machine	because of	its researching a problem more th
your datums with this program .	Because of	its having strong , smart mind yo
teach more lesson in less time .	Because of	getting more time , teachers can
o easy that , you do nt be sorry	because of	absence of a person to play a gam
ay even die without getting help	because of	no way of communication Communica
he past in which the people died	because of	unknown illnesses . The second cl
agriculture . In 20th century ,	because of	the population growth and the gro
people die or have some injuries	because of	the traffic accident . But it is
s . Many people lost their lives	because of	atom bomb . Many had incurable in
ams which we are interested in .	Because of	the advertisements , we know many
rich by paying too high taxes .	Because of	these unfair taxes all the people
ive together instead of marrying	because of	this reason premarital sex is bec

*Figure 17. S*ample concordance of a multiword expression from TICLE

When we compare this list with the one obtained from AntConc using lexical bundle criteria, two of the expressions including *because of the* (58) and *because of this*

(190) exist. Some concordance lines with these expressions may be seen below:

(191) Everybody has some kind of problems, **because of this** problems the people commit a crime. [TRCU1001]

(192) **Because of the** advertisements, we know many things shown [TRCU1021]

(193) Thirdly the families often need two incomes **because of the** difficulties of living in this world. [TRKE2035]

Also, the second expression is *for example* which appeared as *for example in* (27) and *for example the* (23) in our LB list. Furthermore, *have to* expression has been used in structures *I/we/they have to* and *have to be*. Examples are as such:

(194) **For example in** our coutry engineering is tought appropriate [TRCU1116]

(195) **I have to** use both of them effectively in [TRCU1129]

(196) But we do not **have to be** a slave of it ; we ought [TRCU1054]

When looked at the multiword expression in the LOCNESS corpus via the Wmatrix tool, following list has been obtained.

Table 44.

The Most Frequent Multiword Expressions in LOCNESS on Wmatrix

Word	**Semtag**	**Freq.**	**Relative Freq.**
have_to	S6+	113	0.08
such_as	Z5	102	0.07
because_of	A2.2	102	0.07
United_States	Z2	102	0.07
death_penalty	G2.1	78	0.06
not_only	Z5	64	0.05
as_well	N5++	61	0.04
out_of	5	59	0.04
according_to	Z5	55	0.04
public_schools	P1	53	0.04
capital_punishment	G2.1	50	0.04
in_order_to	Z5	49	0.03
their_own	Z8	47	0.03
due_to	A2.2	46	0.03
going_to	T1.1.3	45	0.03
so_that	Z5	38	0.03
used_to	T1.1.1	38	0.03
look_at	X3.4	36	0.03
high_school	P1	35	0.02
a_lot	N5+	32	0.02
as_long_as	Z7	22	0.02
on_the_other_hand	Z4	22	0.02
in_favor_of	S8+	20	0.01
both_sides	N5.1+	20	0.01
based_on	A2.2	20	0.01
whether_or_not	Z7	19	0.01

each_other	Z8	19	0.01
way_of_life	S1.1.1	19	0.01
a_little	A13.6	16	0.01
public_school	P1	15	0.01
human_being	S2	15	0.01
i_think	Z4	13	0.01
up_to	Z5	13	0.01
in_the_past	T1.1.1	13	0.01
a_great_deal	N5+	12	0.01
in_a_way	A13.4	11	0.01
over_the_years	T1.1.1	11	0.01
makes_it	X9.2+	11	0.01
in_this_case	Z4	10	0.01
in_terms_of	Z5	10	0.01
at_the_same_time	T1.1.2	10	0.01
in_some_cases	Z4	10	0.01
more_and_more	A13.3	10	0.01
as_a_whole	N5.1+	10	0.01
a_lot	A13.3	10	0.01
first_of_all	N4	9	0.01
in_my_opinion	Z4	9	0.01
no_matter_what	A7+++	9	0.01
in_front_of	N4	9	0.01
at_the_time	T1.1.1	8	0.01
in_addition_to	N5++	8	0.01
in_the_long_run	T1.1.3	7	0.00
as_it_is	Z4	7	0.00
a_number_of	N5+	7	0.00
in_addition	N5++	7	0.00
most_of_the_time	T1.3+	7	0.00
take_care_of	S8+	7	0.00

***Semtag representing the semantic field numbers**

As the table shows, 57 multiword expressions have been listed by the Wmatrix analysis tool. The frequency cut-off has been 7 and their related semantic tags have been given. The most frequent expressions seem to be *because of, have to, such as, United States, death penalty.* Sample concordance lines from the most frequent item *have to* may be seen below.

atients faced with euthanasia	have to	make a hard decision of whethe
The longing that the parents	have to	fill the biological and emotio
in schools . They should not	have to	take part in a religion that t
cial system so that we do not	have to	" do unto others as they 've d
e women 's game . The viewers	have to	wait until the eleven o'clock
States <*> . Students do not	have to	leave their religion and belie
long run and that they do not	have to	rush into the money . They rea
They realized that they would	have to	order parts for the brakes if
determine this question , we	have to	consider our definition of lif
like it . But other programs	have to	be stopped before they start .
on goes much higher , he well	have to	consider going to a community
Mrs Doe , for example , will	have to	suffer the loss of their child
fford to keep her baby wo n't	have to	kill it , the baby therefore g
ou get a scrape so you do n't	have to	go to the doctor then you are
hing that one would no longer	have to	do themselves . These practice
? The proponents did n't even	have to	analyze ABC 's Play Boy/Play G
out of business , and he may	have to	work long hours for small gain
nals profit (even if they do	have to	wait until after they are rele
hat happens is that all women	have to	go along with these feminists
inflicted . The criminal will	have to	live with his or her actions a
did not think that they would	have to	pay for their crime . It was a

Figure 18. Sample concordance of a multiword expression from LOCNESS

In our list, *have to* have been used frequently within many expressions such as *have to be, not have to, would have to, they have to, have to do, have to go, will have to.* While second expression *such as* appeared as *such as these (7), such as the (16),* the other expression *because of* expression has been used in structures *is because of, because of this, because of the.*

(197) Someone will **have to be** a responsible journalist and say [USARG_0171]

(198) Stories **such as these** certainly put a hold on [USARG_0025]

(199) **Because of this** hacking has become quite a [USARG_0082]

When we compare the multiword expressions in both corpora, the common expressions in both corpora are *each other, according to, a lot of, in order to, on the other hand, in my opinion, first of all, look at, I think, their own, day by day, as well, in addition to, in the past, as it is, in front of, at the same time, more and more, in terms of.* After revealing the multiword expressions with their frequencies in both corpora, key semantic domains of these expressions have been revealed according to UCREL semantic tag set. Semantic domains of the MWIs used both in the TICLE and LOCNESS corpora may be seen below.

Table 45.

Semantic Domains of the LBs in the TICLE and LOCNESS

Rank	Semantic domains in TICLE	Freq.	Semantic domain in LOCNESS	Freq.
1	**Cause & Effect/Connection**	**232**	**Strong obligation or necessity**	**113**
2	**Grammatical bin**	**200**	**Grammatical bin**	**102**
3	**Strong obligation or necessity**	**177**	**Cause & Effect/Connection**	**102**
4	**Pronouns**	**151**	Geographical names	102
5	**Grammatical bin**	**143**	Law and order	78
6	**Quantities: many/much**	**135**	**Grammatical bin**	64
7	**Grammatical bin**	**119**	**Quantities: many/much**	61
8	**Discourse Bin**	**105**	Grammatical bin	59
9	**Education in general**	**88**	Grammatical bin	55
10	Evaluation: Authentic	85	**Education in general**	**53**
11	Evaluation: True	83	Law and order	50
12	**Discourse Bin**	**70**	Grammatical bin	49
13	**Linear order**	**65**	**Pronouns**	47
14	**Sensory: Sight**	**64**	**Cause & Effect/Connection**	46
15	**Discourse Bin**	**63**	Time: Future	45
16	Degree: Boosters	62	Grammatical bin	38
17	**Pronouns**	**62**	Time: Past	38
18	Investigate, examine, test, search	58	Sensory: Sight	36
19	Quantities	57	Education in general	35
20	**Comparing: Usual**	**50**	Quantities: many/much	32
21	**Quantities: many/much**	**44**	If	22
22	Speech acts	41	Discourse Bin	22
23	**Quantities: many/much**	**41**	Helping	20
24	**Time: Past**	**36**	Entire; maximum	20
25	Mental object: Conceptual object	36	Cause&Effect/Connection	20
26	Places	32	If	19
27	Strong obligation or necessity	31	Pronouns	19
28	Evaluation: Authentic	30	Social Actions, States and Processes	19
29	**Discourse Bin**	**27**	Degree: Diminishers	16

30	**Linear order**	**26**	Education in general	15
31	**Time: Present; simultaneous**	**24**	People	15
32	**Degree: Boosters**	**22**	Discourse Bin	13
33	Time: Present; simultaneous	21	Grammatical bin	13
34	Giving	21	Time: Past	13
35	Time period: long	21	Quantities: many/much	12
36	Discourse Bin	21	Degree: Approximators	11
37	Hindering	20	Time: Past	11
38	Discourse Bin	20	Success	11
39	**Grammatical bin**	**18**	Grammatical bin	10
40	Discourse Bin	17	Time: Present; simultaneous	10
41	Discourse Bin	17	Degree: Boosters	10
42	Linear order	13	Entire; maximum	10
43	Discourse Bin	9	Linear order	10
44	Places	6	Discourse Bin	9
45	Residence	6	Likely	9

The common semantic domains in both corpora are highlighted in bold in the Table 46. There exist 12 common categories. The common domains are cause & effect / connection, strong obligation or necessity, pronouns, grammatical bin, quantities: many/much, discourse bin, education in general, linear order, sensory: sight, degree: boosters, time: past, time: present; simultaneous. The most significant domain belongs to the cause & effect/connection among all. Grammatical bins are the second most frequent domain. And strong obligation or necessity and quantities: many/much domains follows respectively in both corpora. Some of the most frequent expressions of these domains are *because of, have to, according to, each other, a lot of, in order to, on the other hand*…etc.

We also used log-likelihood ratio statistics to compare frequencies of the semantic domains of these common multiword expressions in both corpora to see the differences in the usages, which is given in the table below.

Table 46.

The Log-likelihood Results of Semantic Domains in both Corpora

USAS tag	Freq. in TICLE	Freq. in LOCNESS	Log-likelihood	Sig.	Semantic Domain
A2.2	232	102	55.41	0.000 ***+	Cause & Effect/Connection
S6+	177	113	15.96	0.000 ***+	Strong obligation or necessity
Z8	151	19	120.08	0.000 ***+	Pronouns
Z5	143	55	42.85	0.000 ***+	Grammatical bin
N5+	135	32	71.05	0.000 ***+	Quantities: many/much
Z4	105	22	61.16	0.000 ***+	Discourse Bin
P1	88	53	9.72	0.002 **+	Education in general
N4	65	9	49.28	0.000 ***+	Linear order
X3.4	64	36	8.69	0.003 **+	Sensory: Sight
A13.3	62	10	43.16	0.000 ***+	Degree: Boosters
T1.1.1	36	8	20.01	0.000 ***+	Time: Past
T1.1.2	24	10	6.31	0.012 *+	Time: Present; simultaneous

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

According to the table, all semantic domains have been overused in TICLE compared to the LOCNESS. Even though total number of expressions (types in total) used in TICLE are far less than that of expressions in LOCNESS, all the domains produced indicate an overuse in TICLE. The most significant difference (LL value 120.08) in the semantic comparison belongs to the tag *Z8* representing the pronouns. The expressions used with pronouns are *each other, their own.* The second most significant difference is N5+ with 71.05 LL value. This tag represents quantities: many/much domain. Some of the examples of this domain include *a lot, as well, in addition to, a great deal.* The third most significant difference (LL value 61.16) indicates the overuse of the semantic domain of discourse bin (Z4). Examples of discourse domains are *on the other hand, first of all, in my opinion, I think, as it is, in other words*…etc. Another significant category is cause & effect/connection (A2.2) which is also overused in TICLE (LL ratio 55.41). Another overused category is Linear order (N4) with 49.28 LL value. And the rest of the categories are degree: boosters (A13.3 with LL value 43.16) , grammatical bin (Z5 with LL value 42.85), time: past (T1.1.1 with LL value 20.01), strong obligation or

necessity (S6+ with LL value 15.96), education in general (P1 with LL value 9.72) and time: present; simultaneous (T1.1.2 with LL value 6.31) respectively.

It was also found out that there are domains specific to each corpus. The figure below represents the domains specific to each corpus as well as the common ones.

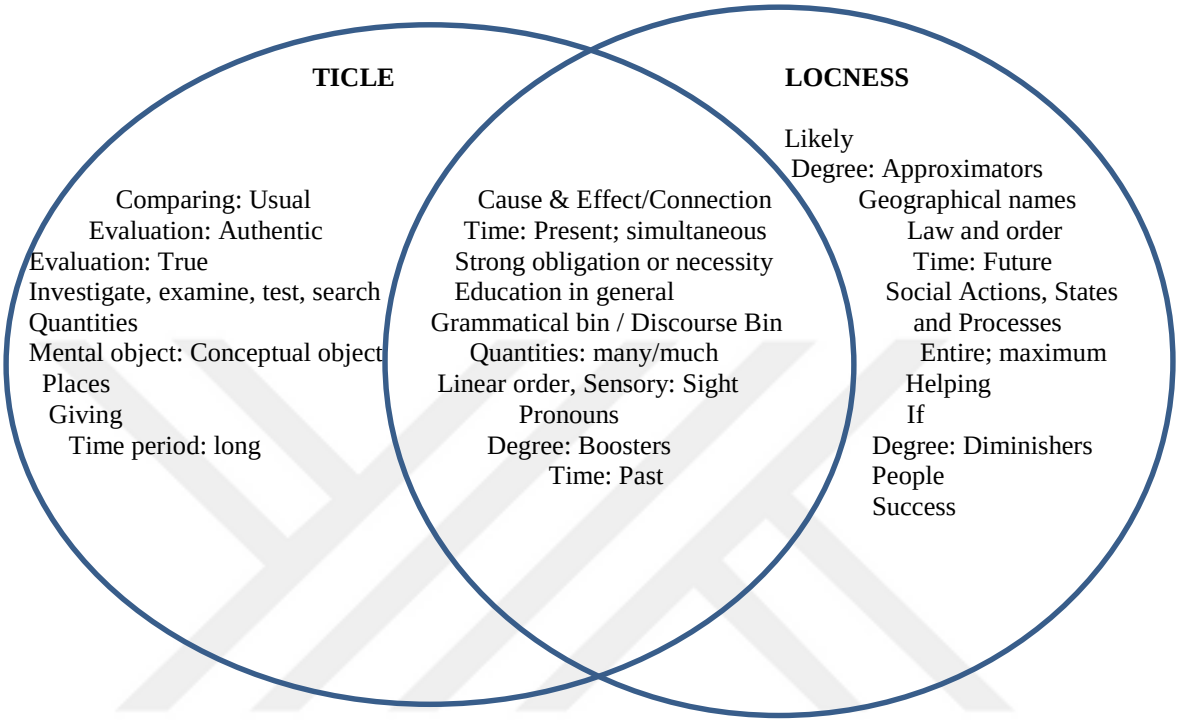


Figure 19. Semantic domains of multiword expressions in two corpora

As the figure represents, while expressions in TICLE have 9 domains that do not exist in LOCNESS which are *evaluation: authentic and true; investigate, examine, test, search; giving; places; time period: long, comparing usual; mental object* and *conceptual object*. Some of the expressions of these domains include *for real, in fact, so many, point of view, all over the world, in other words, day by day, long time*. On the contrary, multiword expressions in LOCNESS have 12 specific domains which are *likely, geographical names, degree: approximators, geographical names, law and order; time: future; social actions, states and processes; entire: maximum; helping; if; degree: diminishers; people* and *success*. Some expressions are *in my opinion, in a way, United States, capital punishment, going to, way of life, both sides, in favor of, whether or not, a little, makes it…etc.*

CHAPTER V

DISCUSSION

5.1. Introduction

In this chapter, the findings of the study will be discussed referring to the previous studies conducted in the field. The results regarding the frequency of 3 and 4-word bundles in TICLE and LOCNESS, structural and functional classifications and comparison of these findings will be discussed in detail. Finally, the summary and review of the research questions are presented.

5.2. Discussion for Research Question 1

RQ1. What are the most frequent 3- and 4-word lexical bundles found in argumentative essays of Turkish EFL learners in the Turkish International Corpus of Learner English (TICLE) and native speakers in the Louvain Corpus of Native English Essays (LOCNESS) corpora?

To address the first research question, the most frequent 3- and 4-word sequences found in argumentative essays by Turkish EFL learners as represented Turkish International Corpus of Learner English (TICLE) learner corpus were extracted. After the extraction of the relevant expressions using AntConc corpus analysis tool (Anthony, 2020) based on both frequency and range criteria, a manual review process was undertaken to expunge expressions that lacked semantic value or were repetitive in nature. During this manual selection phase, careful consideration was given to determine the appropriate frequency cut-off point considering the total number of words in both corpora. The actual frequency cut-offs in lexical bundle research vary from ten occurrences per million words (Biber et al., 1999; Biber, 2006), to forty occurrences (Biber, Conrad, & Cortes, 2004). Lower cut-offs may be used in smaller corpora (Biber et al., 1999) and determining the frequency cut-offs base on the earlier researchers rather than a statistical and empirical set of standards. For this very reason, the cut-off point has been set as 10 for TICLE, 7 for LOCNESS. Lastly, the frequency calculation has been performed to enable comparative analysis of the frequency results across these two corpora.

As a result of the analysis, TICLE corpus was found to comprise of 280 texts with

a word count totaling 199,532. In contrast, the LOCNESS corpus contained 175 texts, yielding a cumulative word count of 149,574. Total number of words used in the study is 349,106 encompassing 455 texts. While Turkish EFL learners employed 340 LBs in their written performances, NS used 349 bundles. Overall number of LBs incorporated into the essays was 689. When we examine the proportional representation of these LBs, only 0.17% of Turkish learners' productions consisted of bundles. As for the native speakers, this proportion slightly increased to 0.23%. In total, 0.19% of these words consist of lexical bundles. In other words, this suggests Turkish EFL learners' comparative underuse of the LBs vis-a-vis the NS. This finding aligns with the findings from previous studies which assert that novice writers use academic-register lexical bundles infrequently as in Chen and Baker's (2010) study. Similarly, native speaker students exhibit a wider diversity in their use of bundles compared to their non-native counterparts as Bychkovska and Lee (2017) also reports.

TICLE corpus has 340 types and 8495 tokens while LOCNESS has 349 types and 4729 tokens. The T/t ratio of the TICLE corpus is 4% in total, LOCNESS corpus has a T/t ratio of 7.43%. This means that 4 of the Turkish learners' every 100 words are LBs. NS used 7.51 of their every 100 words as a bundle. When we compare the two corpora considering their total number of words in both corpora, NS have produced more varied LBs according to the t/t ratio. This suggests that a greater lexical richness and diversity inherent in the native speakers as compared to the corpus of non-native EFL learners. These findings are consistent with Muşlu's (2015) research which focused on the use of stance bundles by Turkish and Japanese EFL learners and NSs. In her study, she concluded that non-native speakers' use of bundles demonstrated significantly less variation.

The number of lexical bundles found in this dissertation is higher than the numbers found in previous studies. This could be attributed to the methodology adopted in our research, which investigates both trigram (3-word) and quadrigram (4-word) expressions. Previous studies, such as Hyland's (2008a) work, revealed 629 bundles in a nearly 3.5-million-word corpus while Wei and Lei (2011) found 241 bundles in a 2.5-million-word corpus. It should be emphasized that these studies focused exclusively on 4-word bundles. Wei and Lei (2011) assert that 3-word bundles are extremely common expressions, and in their analysis of the 100 common expressions, only six of them are 4-word bundles while the rest are 3-word bundles. The quantity of lexical bundles identified in the current study provides substantial insight into the linguistic behavior of Turkish

EFL learners. Supporting our approach, Cortes (2021) suggests exploring frequent 3-word expressions with pedagogical purposes as there are more 3-word bundles than being embedded in 4-word bundles.

Even though TICLE corpus has a total of 340 expressions, LOCNESS corpus slightly exceeds this with 349 expressions. Nonetheless, for the purpose of addressing the first research question, and due to practical considerations, the analysis has been restricted to the 100 most frequent expressions found in the corpora. The pertinent list can be found on page 173-175. Out of 100 most frequent expressions, there exist 40 common expressions in both corpora which are *men and women, a lot of, it is not, in order to, on the other hand, they do not, the most important, in the world, there is no, most of the, there is a, it is a, do not have, it is the, they are not, that it is, there are many, because of the, to have a, should not be+V3, that they are, is not a, the right to, according to the, they want to, in the past, is one of the, to be a, in the future, look at the, this is a, as a result of, part of the, because they are, the fact that, all of the, you do not, do not want, this is not, the number of.* An additional 60 bundles were identified as unique usage instances. This finding aligns with the results of Adel and Erman's (2012) research, which identified 60 shared bundles among a total of 130 within their data set. Conversely, these results are at odds with the findings from Chen and Baker's (2010) study, which reported a higher shared bundle rate, with 54 shared instances out of a total of 78 bundles. They also asserted that ‘‘the use of lexical bundles in nonnative and native student essays is surprisingly similar’’ (p. 44).

When the frequency of these bundles compared, the most frequent expressions by the Turkish EFL learners are *men and women* (f: 126), *a lot of* (f: 123), *it is not* (f: 120), *in order to* (f: 119), *on the other hand* (f: 119), *they do not* (f: 119), *the most important* (f: 109), *in the world* (f: 108), *there is no* (f: 106), *most of the* (f: 98) respectively. The ten most frequently occurring bundles, which amount to 1147 tokens, represent 13% of the total tokens within the TICLE corpus. On the other hand, in LOCNESS corpus, the most frequent expressions are *the death penalty* (f: 71), *the fact that* (f: 71), *one of the* (f: 63), *in the United States* (f: 53) *the right to* (f: 51), *in order to* (f: 50), *because of the* (f: 47), *that it is* (f: 47), *the use of (*f: 45), *it is a* (f: 40). The ten most frequently occurring bundles within the LOCNESS corpus were noted 546 times, constituting 11% of all the tokens in this particular corpus. Some of these most frequent expressions such as *in order to* (f: 119), *one of the* (f: 63), *the use of* (f: 45), *the fact that* (f: 71), *there is a* (f: 86) are among Biber et al.’s (1999) most common 3-word lexical bundles list which include *in*

order to, one of the, part of the, the number of, the presence of, the use of, the fact that, there is a, there is no.

Within these expressions, the nominal phrase *a lot of*, identified as having an informal marker and the phrase structure *the most important* have been discussed as being conversation-type bundles. Using these types of bundles is frequently cited as a distinctive feature of learner writing (Staples et al., 2013; Chen & Baker, 2016; Bychkovska & Lee, 2017). Additionally, Shin (2018) observed a significant presence of these types of lexical bundles in her study. This suggests that the usage of these bundles is not confined to a specific learner population but rather is indicative of novice academic writing, irrespective of the first language. In our study, *a lot of* has been used 123 times by Turkish EFL learners compared to its usage 29 times by native speakers. Similarly, token of *the most important* occurrred 109 times in TICLE corpus, and only 18 times in LOCNESS corpus.

Also, the structures of the bundles extracted from both TICLE and LOCNESS are not complete units in terms of grammatically, as may be seen in the examples above. This finding is in line with the finding of Biber et al.'s (1999) study which found out that more than 95% of the lexical bundles were not complete units in academic writing. In addition, same argument has also been supported by Cortes (2004) stating that "lexical bundles are identified empirically, rather than intuitively, as word combinations that recur most commonly in a register, and therefore, lexical bundles are usually not complete structural units, but rather fragmented phrases or clauses with new fragments embedded" (p. 400).

In addition, when we examined the pronouns used in the lexical bundles in two corpora, it was discovered that Turkish EFL learners overuse pronouns in comparison to native speakers. Specifically, the singular/plural pronoun "you" and plural pronoun "we" were used nearly twice as often in TICLE corpus compared to their occurrence in the LOCNESS corpus. This finding challenges the results of prior studies by Chang & Swales (1999) and Wei & Lei (2011), which suggested that advanced English learners have a tendency to use passive structures and few personal pronouns, thereby maintaining an impersonal tone in their writing. Conversely, the singular pronoun "I" was underused in the TICLE corpus. This phenomenon of the underuse of "I" has also been observed in previous studies by Harwood (2005) and Wei & Lei (2011). These studies corrolate the phenomenon to the instructions provided in learners' preceding English writing courses and advice found in EFL teaching materials, which uphold the notion that formal written language should strive to be objective and impersonal.

5.3. Discussion for Research Question 2

RQ2. How are these lexical sequences found in TICLE and LOCNESS corpora classified based on their structural and functional characteristics?

The objective of this research question was to scrutinize the frequency of lexical bundles in both the TICLE and LOCNESS corpora and to categorize these bundles in alignment with the structural and functional classifications posited by Biber et al. (2004). The structural classification of lexical bundles in the Longman Grammar of Spoken and Written English (Biber et al., 1999) has been extensively applied in other studies focusing on recurrent word combinations (Cortes, 2002, 2004; Hyland, 2008a, 2008b). Nonetheless, to accommodate expressions that did not align with the existing categories, we introduced two novel sub-categories to the original classification schema. The first one is *other adverbial clause fragments* within *the dependent clause fragments* such as *since they are, but it is not, as I said*. The latter is *the quantifier expressions* such as *a lot of, some of them, most of the people* added to *the noun phrase and prepositional phrase* category.

In the course of their argumentative essays, Turkish EFL learners employed 340 distinct bundles, accounting for a total of 8495 tokens. The initial category of the structural classification pertains to verb phrase fragments, represented by 145 types and 3466 tokens, comprising 42.4% of all structures in the TICLE corpus. This echoes the findings of Chen and Baker' (2010) study, which highlighted that student writings typically contain more *VP*-based bundles than those written by native speakers, pointing to trend especially evident in specific subcategories. In line with this observation, our categorization scheme, also identified this trend in the verb phrase fragments subcategory of personal pronouns+VP fragments, which constituted a quarter of all entries in the structural classification. This finding is also in line with Shin's (2018) study which similarly reported that non-native writers employ VP-based bundles more frequently than their native speaker counterparts.

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

preference for negatively phrased expressions such as *disagree with the statement* and *do not agree with*. Conversely, learner writers tended to favor positively phrased expressions like *agree with the statement* or *so I agree with*. Furthermore, in response to the same prompts, the learner writers seldom produced negative examples, whereas native speakers address negative aspects.

A notable observation from the data is the absence of any *discourse marker + VP fragment* such as *I mean you know, you know it was*…etc. and *wh -question fragment* combinations like *what do you think, what does that mean*…etc. within the written productions of Turkish EFL learners. This absence may be attributable to the type of corpus, as these expressions are more common in spoken registers and typically appear in larger clausal expressions with a pronominal subject as per Biber et al.'s (1999) study. The corpora engaged in this research consisted of written argumentative essays. It is also noteworthy that these two structure types absent in El Turki's (2015) study, which examined the development of LBs among learners at various proficiency levels over a one-year period.

Additionally, the analysis uncovered that a substantial portion of the fragments in this subcategory include bundles incorporating embedded *be-verbs* with 44 types, constituting 38.2% of all verb phrase fragments. This finding aligns with the research of Chen and Baker (2016), which reported that a third of the LBs incorporated *be*-verbs. The use of these *be*-verbs can render student writing as "simplistic and verbose" (p. 866). The authors further posited that this overuse of *be*-verbs might be attributed to the learners' extensive reliance on existential *there*-constructions. This pattern was also markedly evident in the argumentative essays of Turkish EFL learners. Examples from our study include *there are some/many/a lot/lots of/many people/ also/so many* (f: 215), *there is a/an* (f: 108), *there is no (f: 89), there is no need* (f: 17), *there is not* (f: 26), *there is nothing* (f: 20). However, in contrast to these findings, Ädel and Erman (2012) indicated that the native students used more existential-*there* constructions and passives, while non-native students tended to initiate arguments with evaluative bundles like *anticipatory it* patterns (e.g., *it is easy to*). Contrary to Ädel and Erman's study, our research discovered that non-native speakers used more existential *there*-bundles at a significantly higher rate than their native counterparts.

With regard to the evaluative bundles such *anticipatory it* patterns as in *it is easy to*, our research aligns with the findings of Muşlu's (2015) study, which reports higher usage of *anticipatory it* patterns by the non-native students (both Japanese L1 and Turkish

L1) than by the native speakers.

Additionally, functional classification analysis revealed that not all the used structures serve as a function in the discourse. The nature of argumentative essays, the specific genre used in our study, could explain this outcome since the primary objective of such essays is "to express [one's] opinion about an issue" (Staples et al., 2013, p. 217). Discourse organizers were most extensively used by both Turkish EFL learners (36%) and native speakers (35%). In the TICLE corpus, stance bundles were the second most commonly used function (35%), while in the LOCNESS corpus, referential bundles filled this spot (34%). Referential bundles were least used by Turkish EFL learners (29%), stance bundles were the least frequently used by the native speakers (28%). There is a slight underuse of the functional properties in the TICLE corpus in comparison to the LOCNESS corpus. However, the overall difference between the two corpora in terms of functional properties is significant with a LL value of 8.19-.

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

Another factor could be the methodology itself, as ''larger corpora will generate fewer recurrent word combinations with the same cut-off normalized frequency, when compared with smaller corpora, because large corpora will elicit higher converted raw frequencies'' (Chen & Baker, 2010, p. 43; see also Biber & Barbieri, 2007, p. 269). While we employed different cut-off points for each corpus, smaller cut-off points might have revealed different bundle patterns.

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

(Shin et al., 2018).

5.4. Discussion for Research Question 3

RQ3. How similar or different are the frequently used 3- to 4- lexical sequencies used by Turkish EFL learners from those found in native English speakers' (LOCNESS) corpora in terms of:

a. structure,

b. function, and

c. diversity in their use?

To address this research question, we conducted log-likelihood (LL) analysis. The initial LL analysis evaluated the difference in type frequencies between the TICLE and LOCNESS corpora to assess any overall frequency disparity. The subsequent LL analysis measured token frequencies of bundles in both corpora.

a. Structure

Looking at the bird-eye view of the LL findings of the structural classification types, it is discernible that two categories are underutilized in TICLE, while verb phrase fragments reveal an insignificant overuse in terms of type frequency. This underuse points to a deficit in lexical variety and richness compared to native speaker language use. Nevertheless, the LL values corresponding to the tokens of these structures indicate an overarching overuse by Turkish EFL learners across all three categories. The overuse of these bundles may be attributed to learners' tendency to go for the safest lexical options labeled as "lexical teddy-bear tendency" by Hasselgren (1994). As a result, Turkish EFL learners underuse other potential alternatives for these existing bundles by clinging to a limited set of phraseological "teddy bears".

A comparison of the structures used in both corpora reveal that while Turkish EFL learners predominantly employed *verb phrase fragments*, native students leaned more towards *noun phrase and prepositional phrase fragments*. This substantiates the evidence presented by Bychkovska and Lee (2017) who found a marked difference in lexical bundle use in argumentative essays by L1 and L2 English undergraduates. According to their findings, learners heavily rely on clausal bundles while the native speakers predominantly use phrasal bundles. This finding also alings with studies by Bal Gezegin (2019) and Wei and Lei (2011), which revealed a strong preference for noun phrase structures among native speakers. Nonetheless, it contrasts with the findings of Shin

(2018), who reported that VP fragments were the most frequently used structures in both native and non-native corpora.

When we examine the sub-categories of these fragments in detail, it is evidently seen that *personal pronouns+verb phrase fragments* constitute the only category that is overused in TICLE in both type and token calculations. These fragments represent a quarter of the entire structural classification. This implies that Turkish EFL learners produce more VP-based bundles relative to native speakers, corroborating the findings of Chen and Baker (2010). Their study also highlighted that student writing typically contains more VP-based bundles than native writing, and this usage is particularly noticeable in certain subcategories. Likewise, this finding also confirms Shin's (2018) study on native and non-native novice writers' use of lexical bundles, which demonstrated that non-native writers use VP-based bundles more frequently than NS.

Within the verb phrase fragments, *discourse markers+VP fragments, wh-question* and *yes/no question fragments* were absent in both corpora. This could be the types of corpora studied, as both corpora consist of written registers made up of argumentative essays. Biber et al. (1999) noted that such expressions are more common in spoken registers.

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

As for the second category, *dependent clause fragments*, a significant discrepancy is observed in terms of *to-clause fragments* which Turkish EFL learners underused substantially. This finding diverges from Chen and Baker's (2010) earlier study, which found that both L1 Chinese student writers and NS used "to-clause fragments" extensively, showing a preference for the frame "in order to + Verb." L1 Chinese students used six different verbs that fit in the slot such as *achieve, avoid, be, maintain, make*, and *understand*, while British students had two bundles similar which are *in order to make* and *in order to minimize.* For this subcategory, they observed more similarity between native speakers and Chinese learners.

Another noteworthy finding pertains to *dependent clause fragments*, where learners frequently use *who-relative clauses* with VP bundles, and LBs containing embedded who-clauses such as *people/students/women who are, the people who are,*

people who have. These structures do not appear in the native speaker corpus. Moreover, most of these structures involve *people*, indicating an overuse of this somewhat vague word, a pattern characteristic of learner writing. Many of the LBs produced include or collocate with *people*. In TICLE, *'people'* is used in 15 bundles, whereas native speakers used 'people' in 8 bundles, suggesting a slight overuse by Turkish EFL learners. This observation aligns with previous studies of Chen & Baker (2016) and Bychkovska & Lee (2017).

The last category, NP and PP structures, indicates that there is a quite significant underuse of noun phrase with *of-phrase* fragments in TICLE in relative to LOCNESS. This result is in line with Shin's (2018) study which resulted that the native writers used significantly more noun phrases with *of-phrase* fragments than their learner counterparts.

Another interesting finding is the overuse of *prepositional phrases* by Turkish EFL learners both in terms of type and token LL values. The common expressions *in the world* used 19 times by the native speakers, while Turkish EFL learners used this expression 109 times. This sub-category features expressions specific to Turkish L1 learners such as *all over the world* (f: 32), *around the world* (f: 12), *all around the world* (f: 18), *in real life* (f: 16), *in the real world* (f: 16), *for real world* (f: 16). Among these expressions, Chen (2009) noted that learners often over-generalize the use *all over the world* classifying it as 'learner bundle' expressions frequently used in L2 academic writing but rarely in native English academic writing (Chen & Baker, 2010). However, this finding contradicts Shin's (2018) study, which found that native speakers used many idiomatic expressions such as *in the long run* (20) and *in the real world* (17) more than the learners.

Further, there is a significant underuse (LL value 7.92-) of other noun phrase expressions in TICLE compared to the LOCNESS. While Turkish EFL learners have mainly used *the real world* (f: 83), *the real life* (f: 41), *point of view* (f: 37), *women and men* (f: 51), *the other people* (f: 27), *his or her* (f: 24), *the students to* (f: 27) in this category, *due to the* (f: 17), *due to the fact* (f: 11), *such as the* (f: 16), *as a whole* (f: 11), *the same as* (f: 8). As can be seen, the bundles used by the NS are more varied. Also, there are some expressions within this sub-category specific to the NS such as *the opposite sex* (f: 15), *a good example* (f: 7), *an attempt to* (f: 7) *church and state* (f: 7), *rules and regulations* (f: 7), *the conflict theory* (f: 7), *the work place* (f: 7), *the united states government* (f: 7), *the American people* (f: 8), *the supreme court* (f: 10), *a strong argument* (f: 10), *the first amendment* (f: 11). This range of expressions might explain the

underuse of other noun phrase expressions by Turkish EFL learners.

b. Function

The functional classification of lexical bundles has followed Biber et al.'s (2004) taxonomy. According to the classification, 227 types of all bundles have been categorized with 5725 tokens in TICLE while out of 349 bundles, only 223 bundles could have been categorized into the functional categories in LOCNESS.

Beginning with the stance bundles used by both groups, the log-likelihood value shows an underuse by Turkish EFL learners in terms of types, but a significant overuse in TICLE in terms of tokens. This somewhat contradicts with prior research indicating that native writers use proportionally more stance bundles than learners in terms of both types and tokens, as suggested by Ping (2009) and Bychkovska & Lee (2017). This also contradicts with Jaworska et al.'s (2015) claim that personal stances are commonly expressed in these types of argumentative essays. Overall, the usage percentages of epistemic devices by native speakers and Turkish EFL learners are fairly similar, with native speakers employing epistemic bundles slightly more. This corroborates Chen's (2012) findings, which showed that non-native (Chinese) speakers used a narrower range of epistemic stances compared to native speakers.

When it comes to using discourse organizers, both Turkish EFL learners and native speakers employ these as their most common function. This aligns with findings from Chen and Baker's (2016) study, which concluded that both British and Chinese students also employed a significant number of discourse organizers in their writing, particularly to elaborate and/or clarify a topic. Most of the structures used are verb phrase-based bundles, such as *'this means that', 'that is to say', 'can be used'*, etc. Another study by Staples et al. (2013) also revealed that in their corpus of EAP texts, more than half of the LBs used functioned as discourse organizers.

As for the referential bundles, these are the least frequently used expressions by Turkish EFL learners while native speakers used them as the second most frequent bundles. Adel and Erman (2012), who revealed that the largest proportion of LBs functioned as referential expressions in their study on LBs in academic writing, attribute the differences between referential expressions and discourse organizers to the frequent use of prepositional phrase-bundles. The minimal use of referential bundles is also observed in previous studies by Shin (2018), Chen & Baker (2010), and Salazar (2014), who found that both native and non-native novice writers used only a few reference

bundles, with natives using slightly more than non-natives. Chen and Baker (2016) also noted that using fewer referential bundles is characteristic of novice writers regardless of their first language. In all these studies, the variations in the proportions of bundles terms of their functions might indicate the different features of different genres. In other words, genre plays an important role in the use of bundles.

Within this category, the most significant difference lies in the underuse of *time/place/text reference* bundles by non-native speakers. This could be due to their avoidance of using prepositions for time, place, and text references. Morris and Cobb (2004) argue that L2 learners often resort to avoidance strategies when dealing with multi-word items. In line with this, Aybek and Can's (2019) study on prepositional verb errors found that Turkish EFL learners avoid using certain prepositions when using multi-word items.

The structural similarities between the two corpora in terms of functional properties are not particularly substantial. This could be attributed to the nature of the (non-disciplinary) argumentative essays, whose primary purpose is to articulate personal opinions on an issue, as stated by Staples et al. (2013, p. 217). Moreover, argumentative essays often embody characteristics of spoken genres, such as personal declarations.

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

c. Diversity in their use

Due to their overuse of lexical bundles, Turkish EFL learners have demonstrated less variety compared to native speakers, often repetitively utilizing a limited number of bundles. This pattern might also be linked to Turkish EFL learners' underutilization of alternative expressions in the target language. This observation aligns with the findings of prior studies like Chen and Baker (2010), which noted that novice writers infrequently use academic-register lexical bundles. Similarly, Bychkovska & Lee (2017) pointed out

that native students have a larger repertoire of bundles than non-native counterparts. Multiple other studies have concluded that non-native writers utilize fewer items compared to native speakers, except for a few high-frequency expressions which they tend to overuse (Cortes, 2004; Granger, 1998; Nesselhauf, 2005).

Assigning functions to the bundles extracted from both corpora was challenging due to the multifunctionality of many lexical bundles. Biber et al. (2004) stress the importance of considering the extended context of bundles in order to determine their functions. Given the larger number of bundles in our study compared to previous research, thoroughly analyzing the context of all bundles and finalizing classifications for both corpora was challenging. Moreover, some bundles could potentially be assigned multiple functions, as Biber et al. (2004) pointed out in their study. Such bundles were examined in detail within their contexts. For instance, bundles like "*the beginning of the*" and *"at the end of*" could function as references to time, place, or text deictic reference. Therefore, all bundles were classified according to their most common uses in our study.

The functional discrepancies between our study and previous ones may stem from several factors. Firstly, non-native writers seem to over-generalize and prefer certain idiomatic expressions known to be specific to L2 writing. As suggested by Laufer & Waldman (2011), for English learners to use written and spoken English more fluently and effectively, it is crucial to acquire and practice multi-word items.

Another factor may relate to learners' avoidance strategies. "Avoidance" refers to situations where L1 speakers evade the use of certain L2 structures they find challenging (Alonso, 2002). This pattern was evident in Aybek and Can's (2019) study, which found that Turkish EFL learners tend to avoid using certain prepositions when dealing with multi-word items. In our study, learners repeatedly used the same structures instead of potential alternatives. Using dependent clause fragments as the least used structures, Turkish EFL learners may have, again, avoided using relative clauses or other structures formed with dependent clauses in their writings. In addition to avoidance, learners may have also used missing or unnecessary items in certain structures, or errors may have gone unnoticed during analysis. We also excluded erroneous structures during the manual phrase selection stage. Additionally, fossilization, defined as persistent non-native structures in the learner's interlanguage or inability of a learner to produce native-like performance, might explain the repetitive use of the same structures (Aybek and Can, 2019). For some learners, fossilization could be a factor. The underuse of the aforementioned categories may also be due to a lack of knowledge about the rules. As

Keshavarz (2012) suggests, errors arising from this lack reflect the learner's proficiency at a specific stage of second language development and illustrate some general characteristics of language learning.

The differences with other studies might also stem from learners' reliance on their native language (L1 transfer), where they "choose words and phrases closely resembling their first language or those learnt early or widely used" (Hasselgren, 1994, p. 237). Furthermore, cultural writing styles or habits could also influence the writers.

5.5. Discussion for Research Question 4

RQ4. What are the semantic domains of multi-word expressions in both TICLE and LOCNESS corpora? How similar and different are they from each other?

The UCREL Semantic Analysis System (USAS) and the Wmatrix (Version 4) semantic analysis platform (Rayson, 2008) were utilized to identify the most prominent features of learners' multiword choices. This approach was taken because, despite previous studies exploring the functions of bundles in discourse, very few studies have linked these lexical bundles (LBs) to their semantic domains, as discussed by Cortes & Hardy (2013). The most frequent LBs were selected from both corpora, and their semantic domains were revealed and compared.

Between the two corpora, there were 12 common domains: *cause & effect / connection, strong obligation or necessity, pronouns, grammatical bin, quantities: many/much, discourse bin, education in general, linear order, sensory: sight, degree: boosters, time: past, time: present; simultaneous*. The most frequent domains belong to *the cause & effect/connection, grammatical bin, strong obligation or necessity, quantities: many/much* in both corpora. The *cause-and-effect connectors* and *grammatical bins* include the conjunctions, connectors, adverbials and auxiliary verbs used, which are naturally common structures. *Strong obligation* or *necessity* include different modals used in the essays. The inclusion of *time domains* in both corpora is understandable as learners preferred using all tenses in their essays.

One shared domain, education in general, focuses on education as a semantic category. This is expected since both native and non-native speakers are university students and some of the essay topics are related to education in the corpora. This semantic domain has been frequently observed in Shin's (2018) study of LBs in L1-Korean-speaking EFL learners and native novice writers.

When comparing the frequencies of these semantic domains in both corpora using log-likelihood ratio statistics, it was found that Turkish EFL learners have overused all the domains compared to native speakers. The most significant difference (LL value 120.08) in the semantic comparison belongs to the tag Z8 representing pronouns. This aligns with the findings (RQ1) that non-native students have overused pronouns compared with native speakers. This lesser use of pronouns by native speakers is also observed in other studies such as Chang & Swales (1999) and Wei & Lei (2011), which found that native speakers tend to use passive structures and fewer personal pronouns in their written productions.

The second most significant difference is represented by the tag N5+ (Quantities: many/much) with a log-likelihood value of 71.05. This tag represents *quantities: many/much* domain, with examples including phrases like *a lot, as well, in addition to, a great deal*. The third most significant difference (LL value 61.16) indicates the overuse of the semantic domain of *discourse bin* (Z4). Examples of discourse domains are *on the other hand, first of all, in my opinion, I think, as it is, in other words*…etc. These overuses have also been revealed in the RQ2 and RQ3 above.

The Turkish International Corpus of Learner English (TICLE) exhibits 9 specific domains which are not present in the Louvain Corpus of Native English Essays (LOCNESS), including *evaluation*: *authentic and true; investigate, examine, test, search; giving; places; time period: long, comparing usual; mental and conceptual object*. Turkish EFL learners frequently express their opinions using bundles like *in my opinion* and *point of view* which fall within the domains of *evaluation: authentic and true and mental and conceptual objects*. Additionally, learners produced time- and place-related prepositions more frequently than their native counterparts. These findings are expected since the essays are primarily opinion pieces requiring these types of expressions.

Conversely, LOCNESS contains 12 specific domains that are not present in TICLE. These include *likely, geographical names, degree: approximators, law and order; time: future; social actions, states, and processes; entire: maximum; helping; if; degree: diminishers; people; people and success*. Native speakers exhibit a richer and more varied use of these semantic domains. For instance, native speakers frequently use the phrase *United States* in their essays to refer to the country they live in, the culture they have, or to the people who come and go and live there while this is not the case in non-native speakers' productions. *The system of law and order* is often argued by NS, as well. Also, degree: approximators relates to the time expressions used. While NS focus on the

present and future generally, Turkish learners do not mention about the future as frequent as the native speakers do. All these domains specific to NS reveal the lexical variety and richness in relative to Turkish speakers. To exemplify, NS preferred using *whether or not*, or *as long as* phrases instead of using *if*. However, this is not the issue in TICLE. For these reasons, these domains may have been observed in LOCNESS corpus.

Rayson (2008) proposes that looking at the semantic domains assists interpreting qualitative examination. It could also serve as a useful mean to compare writing produced by different language groups, and further provide immediate implications for different populations of writers. The reasons behind these difference between two corpora and with the other studies are not quite clear and require further investigation. It should also be noted that the topics for the assignments varied to a great extent in these two student corpora even though both corpora are comprised of argumentative essays. Most of the essays' topics relates these issues, covering many fields in general. That may be one of the reasons for these differences in both corpora. Another reason may be the differences between the cultures and point of views between learner groups.

CHAPTER VI

CONCLUSION AND SUGGESTIONS

6.1. Introduction

This final chapter wraps up this dissertation by outlining the study's findings (Section 6.2), proposing directions for future research, and discussing the limitations of the investigation. Pedagogical implications will also be provided for EFL teachers, researchers, instructors working at ELT departments as well as publishers.

6.2. Conclusion

Multiword units that "are important building blocks of discourse in spoken and written registers" (Biber & Barbieri, 2007, p. 263) are also referred to as formulaic language/sequences by some scholars (Conklin & Schmitt, 2008; Schmitt, 2004; Wray, 2000). These phrases are crucial for second language competence as they enhance the fluency and quality of both oral and written communication. Acquiring these multiword items helps learners become proficient and fluent (Thornbury, 2002; Boers et al., 2006).

Frequency is a key consideration for these multiword items as it indicates that a sequence of words is stored and utilized as a pre-formed chunk. Sequences with higher frequency are more likely to be stored as unanalyzed chunks than those with lower frequency (Biber et al., 2004). Almost every theoretical description of the corpus-driven approach mentions frequency, such as the idea that "linguistic categories" are derived "systematically from the recurring patterns and frequency distributions that emerge from language in context" (Tognini-Bonelli, 2001, p. 87), or in a grammatical pattern, "a combination of words occurs relatively frequently" (Hunston & Francis, 2000, p. 37).

Lexical bundles, defined as "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" by Biber et al. (1999, p. 990), represent a significant type of MWIs that are prevalent, and constitute an important portion of discourse. Biber et al. (1999) noted that in their academic prose corpus, 21% of all the words were recurrent lexical bundles. In addition to their recurring nature, lexical bundles possess distinctive attributes that differentiate them from other types of multi-word expressions such as collocations and idioms. Some examples of lexical bundles are kinds of expressions such as *as a result of, take a look at, on the other hand,* and *I don't know, to be able to, are you going to, I don't know if, at the same time, have*

a lot of, you know it was. These entities can be discerned using computer software that scans for clusters of three or more words frequently co-occurring in textual language. In contrast to other types of word combinations that occur, lexical bundles are detected empirically rather than perceptually within a corpus. They represent the most frequent multi-word sequences in a corpus, serving specific functions in discourse.

This thesis explores the usage of 3- and 4-word lexical bundles in the Turkish International Corpus of Learner English (TICLE), a learner corpus comprising written argumentative essays by Turkish EFL learners. The goal is to illuminate the use and variations of formulaic sequences among Turkish EFL learners compared to native speakers (NS). The Louvain Corpus of Native English Essays (LOCNESS), a native reference corpus containing American students' written essays (Granger et al., 2009), serves as the native corpus used in this study.

This study used the AntConc corpus analysis tool to identify items. For the Turkish International Corpus of Learner English (TICLE), a cut-off point of 10 was determined, while for the Louvain Corpus of Native English Essays (LOCNESS), a cut-off point of 7 was used. Multi-word items appearing in at least 3 distinct texts were selected based on range criteria. Following identification, lexical bundles in both corpora were classified according to structural and functional characteristics, using the framework by Biber et al. (1999; 2003; 2004). This structural classification is widely utilized in similar studies (Cortes, 2002, 2004; Hyland, 2008a, 2008b). Log Likelihood (LL) analysis was then used to highlight similarities and differences between Turkish EFL learners and native English speakers. Finally, the semantic domains of multi-word expressions in both corpora were identified using the UCREL Semantic Analysis System (USAS) and Wmatrix (Version 4) (Rayson, 2008).

The TICLE corpus consists of 280 texts totaling 199,532 words, while the LOCNESS corpus comprises 175 texts with 149,574 words. In total, the study used 455 texts and 349,106 words. Turkish EFL learners used 340 multi-word items, compared to 349 by native speakers, totaling 689. Only 0.17% of Turkish learners' texts consisted of bundles, compared to 0.23% for native speakers. In other words, Turkish EFL learners used fewer lexical bundles than native speakers. This is consistent with previous studies suggesting novice writers use lexical bundles less frequently (Chen and Baker, 2010; Bychkovska and Lee, 2017).

The TICLE corpus contained 340 types and 8495 tokens, while the LOCNESS corpus contained 349 types and 4729 tokens. The type-token ratio for the TICLE corpus

was 4%, and for the LOCNESS corpus, it was 7.43%. This indicates that native speakers produced a more varied range of multi-word items. Hence, the lexical richness and diversity in the native corpus exceed those of the non-native EFL learners.

Looking at the most frequent items in TICLE, bundles such as *a lot of, the most important, in the world* are considered to be conversation-type bundles and using these items is unique to learner writing as many researchers including Staples et al. (2013), Chen & Baker (2016), Bychkovska & Lee (2017) discuss. Out of the 100 most frequent bundles in both corpora, there exist 40 common bundles which are *men and women, a lot of, it is not, in order to, on the other hand, they do not, the most important, in the world, there is no, most of the, there is a, it is a, do not have, it is the, they are not, that it is, there are many, because of the, to have a, should not be+V3, that they are, is not a, the right to, according to the, they want to, in the past, is one of the, to be a, in the future, look at the, this is a, as a result of, part of the, because they are, the fact that, all of the, you do not, do not want, this is not, the number of.* The other 60 bundles are unique usages.

The structural classification analysis of both corpora revealed differences in the structures most frequently used by non-native speakers (NNS) and native speakers (NS). NNS predominantly used verb phrase fragments, accounting for 42% of structures, whereas noun phrase (NP) and prepositional phrase (PP) fragments were the most common phrases used by NS, constituting 46% of the structural classification. In the TICLE, NP and PP fragments were the second most common category, making up 33% of the classification. Meanwhile, in the Louvain Corpus of Native English Essays (LOCNESS), VP fragments were the second most frequent structures, representing 33% of all bundles. Dependent clause fragments were the least common in both corpora. Reason for using DC fragments as the least used fragments may be because of the fact that native speakers may have produced the structures by reducing related relative clauses in their essays. Reduced clauses were not apparent in the analysis. As for the Turkish EFL learners, it is more likely that they may have avoided using these expressions, especially with relative clauses, as they are at the intermediate proficiency level.

The analysis revealed that NNS prefer to use clausal bundles, while NS mostly opted for phrasal bundles. Interestingly, Turkish EFL learners used VP fragments with personal pronouns significantly more than NS, despite Turkish being a pro-drop language. This might be because Turkish EFL learners avoid using passive structures and NP fragments in their writing. Alternatively, it is also possible that they used passive

structures or NP fragments that went unnoticed, as erroneous structures were not included in this study.

Within the VP fragments, it was also revealed that large number of the phrases that Turkish EFL learners used report negative states which can be seen in the examples of *students/universities/people/I/you/they/we/it do(es) not* (f: 214), *they/we/people are not* (f: 90), *they/it should not be* (f: 25), *they should not* (f: 11), *I/they/you/we/he/women cannot* (f: 194)…etc. The analysis found that negative structures make up 22.6% of the category that includes personal pronouns and verb phrase fragments. However, this trend is not observed in the structures used by native speakers.

In addition, the Turkish International Corpus of Learner English revealed that many fragments within the personal pronouns and VP fragments category include bundles with embedded *be* verbs. This overuse of *be* verbs may be related to learners' reliance on *there* constructions, which express *existence*. Indeed, it has been observed that Turkish EFL learners also heavily rely on these existential *there* constructions. The examples in our study include *there are some/many/a lot/lots of/many people/ also/so many* (f: 215), *there is a/an* (f: 108), *there is no (*f: 89), *there is no need* (f: 17), *there is not* (f: 26), *there is nothing* (f: 20). The patterns observed in TICLE are not mirrored in LOCNESS. The analysis revealed that non-native speakers used a significantly higher percentage of existential *there* bundles than their native counterparts. Additionally, the research found that Turkish L1 learners used more evaluative bundles, such as anticipatory *it patterns* like *it is easy to, it is a fact*… etc. than native speakers.

As for the *"discourse marker + VP fragment"* such as *I mean you, you know it was and I mean I don't* and *"WH-question fragments"* including *what do you think, how many of you, what does that mean…etc*. have not been produced by Turkish EFL learners at all. Similarly, NS have not used *"discourse marker + VP fragment"* and *yes/no question fragments*. The research identified only a single instance of *wh-question* fragments used by Turkish EFL learners throughout the data. Interestingly, this usage mirrored that of native speakers, potentially due to the nature of the corpora being studied. Both corpora comprise argumentative essays, and such expressions are typically more common in spoken registers.

Regarding verb phrases with *passive verbs*, no significant differences were noted between TICLE and LOCNESS, despite a slight underuse in TICLE. This aligns with existing literature, suggesting that native speakers employ more passive structures than non-native speakers. Interestingly, all passive structures used by Turkish EFL learners

were formed with modal verbs. It is also possible they may have avoided using passive structures instead of *verb phrase fragments* due to difficulties forming these structures, or because of the potential for errors.

In the second category, dependent clause fragments, a notable difference was observed with respect to *to-clause* fragments, which Turkish EFL learners significantly underused. Furthermore, the learners frequently used *who*-relative clauses with verb phrase bundles and lexical bundles containing embedded *who*-clauses, such as *people/students/women who are, the people who are, people who have*. Such structures were not found in the native corpus. Additionally, many of these structures were formed with the term *people*, a vague word which is often indicative of learner writing. A substantial number of lexical bundles produced in TICLE included or collocated with the word *people*, with a slight overuse by Turkish EFL learners compared to LOCNESS.

The last category, *noun phrase and prepositional phrase* structures, is the most frequent structures used by the native speakers. Turkish EFL learners significantly underused *noun phrase with of-phrase fragments* in relative to native speakers. Another finding within this category is the use of *prepositional phrases* by Turkish EFL learners who overused these phrases than their native counterparts in terms of both type and token LL values. Some of the expressions are specific to Turkish L1 learners within this sub-category such as *all over the world* (f: 32), *around the world* (f: 12), *all around the world* (f: 18), *in real life* (f: 16), *in the real world* (f: 16), *for real world* (f: 16). Likewise, among these expressions, one of the common expressions *in the world* is extremely used by Turkish EFL learners. These expressions have been labelled as "learner bundles" in the related literature. Additionally, another significant underuse is *other noun phrase expressions* in TICLE compared to the LOCNESS. While Turkish EFL learners have mostly used *the real world* (f: 83), *the real life* (f: 41), *point of view* (f: 37), *women and men* (f: 51), *the other people* (f: 27), *his or her* (f: 24), *the students to* (f: 27) in this category, native speakers produced the expressions of *due to the* (f: 17), *due to the fact* (f: 11), *such as the* (f: 16), *as a whole* (f: 11), *the same as* (f: 8). As can be seen, the bundles used by the native speakers are more varied. Also, there are some expressions within this sub-category specific to the NS such as the *opposite sex* (f: 15), *a good example* (f: 7), *an attempt to* (f: 7) *church and state* (f: 7), *rules and regulations* (f: 7), *the conflict theory* (f: 7), *the work place* (f: 7), *the united states government* (f: 7), *the American people* (f: 8), *the supreme court* (f: 10), *a strong argument* (f: 10), *the first amendment* (f: 11). This could be attributed to the differences in cultural features and

society regulations.

Overall, the LL findings of the structural classification types indicate that the only overused fragments are *VP fragments* in TICLE. Specifically *personal pronouns+VP fragments* constitute almost half of the bundles used. Underusing number of types in TICLE indicates that lexical variety as well as richness is too less than that of LOCNESS. However, the LL values of tokens of these structures indicate a complete overuse by Turkish EFL learners across all three categories, indicating that few numbers of bundles have been used numerous times by Turkish EFL students repetitively. Using a limited set of phraseological teddy bears, Turkish EFL learners may not have produced other possible or different alternatives for those existing bundles with the concerns of making errors and/or being inclined to use what they already know.

On the other hand, functional classification of both corpora revealed that 67% of bundles in TICLE and 64% of the bundles in LOCNESS did not have any functions. More than half of the bundles used did not have any functions in the discourse such as *is not the, the use of, aware of the, is not enough, the importance of, the right to, to have a, is not a, the general public, the power to, to take the, to stay home, a way of, an increase in, amount of money, a member of, is not to, the ability to, the work place, the history of, the minds of, the presence of, rules and regulations…etc.*

The classification also indicated that both Turkish EFL learners (36%) and native speakers (35%) have used discourse organizers most. They produced discourse organizers specially to elaborate and/or clarify a topic. Most of the structures used are VP-based bundles. Examples are *this means that, that is to say, can be used*…etc. While stance bundles have been used as the second most frequent function in TICLE (35%), referential bundles are the second most frequent function in LOCNESS (34%). And the last category used by Turkish EFL learners is referential bundles (29%), whereas the stance bundles are the least used function by the native speakers (28%). LL values of functional categories indicate an underuse by Turkish EFL learners in terms of types, however, there is a huge overuse in TICLE in terms of the tokens which indicates that Turkish EFL learners have underutilized all three functional categories. The total discrepancy yields a negative LL value of 8.19, indicating a highly significant difference. In simpler terms, when it comes to utilizing functional bundles, the lexical diversity and richness found in the TICLE corpus are not as extensive as those found in the native speaker corpus. Token counts reveal a significant overuse of all categories in the functional taxonomy by Turkish EFL learners, excluding the identification/focus sub-category. It is apparent that the

higher token count in TICLE is primarily due to overuse of a few identical bundles. By overusing the same bundles, Turkish EFL learners generated less diverse bundles compared to native English speakers and repeatedly used a limited set of bundles. This can also be attributed to Turkish EFL learners underutilizing alternative expressions in the English language.

The differences may be due to the nature and impact of the (non-disciplinary) argumentative essays since the argumentative essays require students to express their own ideas and opinions and these essays mostly consist of spoken genre features like personal statements which are frequently expressed in these types of tasks. Besides, the essays had been compiled under exam situations in both corpora. For these reasons, some of the sub-categories of both structural and functional classifications as well as some structures may be absent in both corpora. These factors may have influenced the results and findings.

In summary, while our research reveals certain common usages of lexical bundles, it also demonstrates that both native and non-native English speakers also favor some specific bundles. These include a pronounced use of *verb phrase-based* bundles, an overuse of *existential-there* constructions, an underuse of *passive structures*, a significant overuse of the same bundles, and an overgeneralization of some bundles. These characteristics align with features of learner writing, as described in relevant literature. Ultimately, we found that lexical diversity and richness in TICLE fall short when compared to native English speakers. When examining the tokens of the identical bundles, there's a marked overuse by Turkish EFL learners, indicating that these students repeatedly produced the same bundles. This may relate to their avoidance of using different and alternative structures of learners or their tendency to use the structures they are already familiar with.

As for the last research question, UCREL Semantic Analysis System (USAS) semantic tag sets have been used utilizing Wmatrix (Version 4) semantic analysis platform (Rayson, 2008) to identify the most salient features of the multiword choices of learners. Between two corpora, there exist 12 common domains which are *cause & effect / connection, strong obligation or necessity, pronouns, grammatical bin, quantities: many/much, discourse bin, education in general, linear order, sensory: sight, degree: boosters, time: past, time: present; simultaneous*. The most frequent domains belong to *the cause & effect/connection, grammatical bin, strong obligation or necessity, quantities: many/much* in both corpora. The *cause-and-effect connectors* and *grammatical bins* consist of the *conjunctions, connectors, adverbials* as well as *auxiliary*

verbs used. These may appear as the most common structures naturally. *Strong obligation* or *necessity* include different modals used in the essays. As for the frequency of *time domain* in both corpora, it is also understandable since both learners preferred using almost all tenses in their essays.

In summary, when comparing the frequencies of semantic domains in both the Turkish International Corpus of Learner English (TICLE) and the Louvain Corpus of Native English Essays (LOCNESS), using log-likelihood ratio statistics, Turkish EFL learners have shown an overuse of all domains in relation to native English speakers. The most significant difference, with an LL value of 120.08, is in the semantic comparison associated with the tag Z8, which represents *pronouns*. This aligns with the findings of our first research question, indicating an overuse of pronouns by non-native students compared to native English speakers. The TICLE corpus has nine specific domains that are not present in LOCNESS, while the LOCNESS corpus has 12 specific domains absent in TICLE. An analysis of the semantic domains reveals that native speakers generated more varied and lexically rich expressions compared to Turkish EFL learners.

These observed differences between two corpora could be due to various reasons. These reasons require further investigation. It should also be noted that the topics for the assignments varied to a great extent in these two student corpora even though both corpora are comprised of argumentative essays. Another significant factor could be the broad range of topics covered in the argumentative essays from both corpora.

Overusing the same bundles reveal that Turkish EFL learners seem to over-generalize and favor certain expressions. Most of the studies also revealed the overusing same bundles is the specific to non-native learner writing regardless of their different L1 backgrounds. Turkish EFL learners may have also avoided using alternative expressions instead of the existing ones. Learners avoid using new expressions when they hesitate making errors. Avoidance could be one of the best explanations of underusing various items in TICLE. In addition to avoidance strategies, learners may have also used missing or unnecessary items in certain structures or erroneous usages may have been unnoticed at the stage of analysis as we excluded the erroneous and repetitive structures at the stage of manual selection of the phrases. Although erroneous uses are frequent in learner writing, we chose not to include them in our study. Studies which included the erroneous structures revealed that learners' attempts to use different bundles are significant and results change when these mentioned structures are included in the analysis. The results would have differed had we not excluded the erroneous and repetitive structures from the

analysis.

Another reason of underusing might be the lack of the knowledge of the structures. Turkish EFL learners may not have simply known the required or alternative structures and their various functions in the discourse. Underusing items stemming from the lack of the knowledge of the rules reflect the learner's competence at a particular stage of second language development as well as illustrate some of the general characteristics of language learning. Furthermore, the Turkish EFL learners may have also transferred their native language (L1 transfer) by producing words or phrases resembling their first language.

Last but not least, cultural writing styles or cultural writing habits may have also had an influence on the writers since the results differ across different L1 backgrounds. The differences between the cultures and cultural writing styles as well as point of views between learner groups should be taken into account.

6.3. Implications for English Language Teaching as a Foreign Language

The results of this study provide valuable pedagogical insights for English Language Teaching as a Foreign Language. Studies in the field particularly underline that the knowledge of multiword sequences is strongly correlated with language proficiency levels (Staehr, 2009; Hyland, 2008a). It has also been noted that foreign language learners fail to process the formulaic language in the same way as native speakers do. Ellis et al. (2008) explains this by native speakers' opportunity to encounter even with low-frequency sequences while learners lack such an opportunity. Hyland (2008a) also states that formulaic language is difficult to acquire and that failing to use it correctly identifies learners as outsiders.

In this study, we have also revealed that Turkish EFL learners either overuse or underuse certain lexical bundles in comparison to native speakers (Section 4.3 and 4.4). This probably implies some underlying problems in EFL teaching pedagogy in Turkish context. The learner corpus-based studies on the use of multiword items, as in this particular study, contributes to building learner profiles of the learners coming from various L1 backgrounds. The profiles of these learners offer significant insights that can guide the development of customized language teaching materials in ELT. Moreover, these profiles also inspire EFL teachers to design bespoke educational resources and corrective teaching activities, techniques, and methodologies.

Our study also implies that MWIs need be taught explicitly, and students should be made aware of the differences and the importance of these expressions. The observed

patterns of overuse and underuse of expressions clearly provide clear guidance on teaching methods for contextual usage of lexical bundles, taking into account different genres. Incorporating multi-word expressions into their writing can pose a challenge for novice student writers and English learners alike.

There exists a great potential for developing ways to facilitate the acquisition of these multiword items. Automatization of formulaic sequences is crucial for acquiring these items (Wood, 2015). Automatization may be promoted through different sets of *preparation* (Ellis, 2005), *focused practice* (Nation, 1989), and *repetition* (Wood and Namba, 2013) tasks. Besides, *focus on form* (FonF) (Long, 1988) is another significant element for the acquisition of the multiword items by directing learners' attention to language forms and structures while they are also engaged in focusing on meaning.

One of the ways to facilitate acquiring formulaic language is *chunking of text* (Wood, 2015). In this teaching technique, sequences are highlighted and checked by means of online/offline tools or sources. Using this technique as a key in their study, Boers et al.'s (2006) learners began to use more multi-word items in their writings in relative to the control group.

Furthermore, *fostering learner awareness* about the use of multiword items in their writings could serve as another effective strategy in teaching these constructs. The use of learner data in the classroom exercises such as comparing learner and native speaker data and analyzing errors in learner language has been suggested to raise awareness (Meunier, 2002). Activities such *as eliciting the collocations, completing collocations from memory, matching the words together, selecting the missing word* to teach collocations and MWIs are necessary for learners to acquire the required language items. All these help students become aware of gaps between their interlanguage and the language they are learning.

Another technique used in awareness raising activities is *typographic enhancement and glossing* (Boers et. al, 2016), in which multiword items are highlighted in texts and explained between lines. In his study, Peters (2012) also found out that learners could recall and produce more sequences after using this technique in the classroom.

Flooding the input is another way to facilitate the acquisition of such sequences. this approach involves creating and abridging texts so that multi-word items appear frequently in discourse. Webb, Newton and Chang (2013) abridged a graded reader so that target collocations could appear a number of times. Four different groups of

intermediate ESL learners engaged with one version through reading and/or listening. Post test results indicated that increased exposure to a collocation generated more production, supporting the notion of the power of frequent exposure.

Last but not least, *the use of dictionary resources and corpus information* in teaching the formulaic language proved to be an invaluable source in teaching English language. Laufer (2011) carried out a study on collocations, that involved giving learners a handout for a missing word activity and allowing them to reference dictionary resources to verify the expressions. Post test of the study revealed a 40% success rate in the use of the collocation. In another study by Wu, Witten, and Franken (2010), a corpus of multiword items was compiled, and learners were asked to use this corpus to find the collocations of the particular words. At the end of the study, learners were able to employ collocations successfully with a 67% success rate.

When learners refer to the concordance lines to see whether the structures they use are correct, concordances can reinforce the learning process (Chambers, 2019). By providing many authentic language samples, learners notice the various uses of the language as also proposed by Schmidt's (1990) '*noticing'* hypothesis which posits that learner needs to encounter a particular language feature multiple times before they can effectively employ it. Those encounters are referred to as *frequency* in a language learning context. Besides, they can also compare their productions with that of native speakers, consult a learner corpus, and correct characteristic interlanguage issues such as underuse, overuse, misuse, all of which can enhance their writing skills. *Explicit instruction* on features specific to writing as well as essay-organizing structures would be beneficial for writers at all levels. In this way, learners may better understand the patterns and usages and integrate those patterns into their own productions. Additionally, the findings revealed that learners heavily rely on *be*-verbs and existential *there*-constructions. They could gain significant benefits from *form-focused instruction* (Ellis, 2012) by exploring examples with LBs embedded within context and focusing on structures that commonly co-occur with different types of MWIs as well as their frequency across different registers.

6.4. Suggestions for English Language Teaching as a Foreign Language

The insights from this study could provide immense value to a diverse range of roles such as teachers of beginner-level academic writers, instructors, foreign language teaching classroom educators, researchers in applied corpus linguistics, and publishing

professionals. The practical applications of these insights could greatly influence methods of teaching and research in these fields.

It is obvious that language learners should be exposed to authentic data wherever and whenever possible. It is extremely necessary for English learners to acquire and practice multi-word items in order to use written and spoken English in a more fluent and effective manner. Given the abundant knowledge about the roles of multiword items in acquisition and production, it is crucial that these formulas should be taught more directly in the classroom. There are also numerous ways these items may be integrated into the EFL instruction/curricula and teachers can devise lessons with a specific focus on formulaic language as mentioned above.

To start with, the direct and indirect applications of corpora and its potential in language pedagogy have been explained in detail on page 13-14. Among those, many cognitive, pedagogical, and motivational benefits of DDL have been revealed by many researchers such as O'Sullivan (2007), Boulton (2009; 2010), Granath (2009). The results of these kind of studies should be integrated into ESL/EFL curriculums. Syllabus design can benefit from the use of specialized corpora in language teaching, tailored to students' proficiency levels and specific needs. Language teachers need to be enlighted about the utilization of corpus data and corpus-based teaching methods within their classrooms This can help students gain a deeper understanding of these structures and assist them in sounding more like native speakers. The incorporation of formulaic sequences into the current syllabus should be promoted, and learners should be encouraged to verify their own language usage by referencing existing corpora. Can (2016; 2017) also emphasizes that EFL teachers should be equipped with the skills of applying data-driven and corpus-based teaching techniques to use in their classrooms.

Corpora should be adopted by foreign language teachers and learners in their classes along with data-driven materials which can easily be developed using a corpus. To exemplify, especially English for academic purposes (EAP) may benefit from Coxhead's (2000) Academic Word List (AWL), a corpus-based product to be use for teaching academic writing. It would be the most relevant and useful decision for EAP learners as it consists of the vocabulary items that are the most frequent and common. In a similar vein, Simpson-Vlach and Ellis (2010) devised an Academic Formulas List (AFL), which includes word combinations occurring significantly more often in academic than in non-academic writing, by taking Coxhead's word list idea to phrase level. Corpora also need to be integrated into teacher education at ELT departments. However, it is still

a neglected area in language teacher education. Instructors as well as prospective teachers at the ELT department should be made aware of the corpora and its usage since they do not know what corpus is and how to benefit from it as revealed in previous studies.

There exist many specialized corpus types that could be used in the classrooms or by the learners themselves. One way to realize the integration of the corpora applications into the English language teaching is using the freely available and user-friendly web-based tools that have become available such as English-corpora.org (formally BYU: https://www.english-corpora.org/) (Davies, 2008-), Lancaster University BNClab (Gablasova, 2020), BNC (2014), and Corpus Mate (https://corpusmate.com, Crosthwaite, 2023). These software let students discover the language by themselves, see the words in their contexts, expose authentic language data, enabling them check their errors and use them while doing their homework. Teacher can make use of these tools in the classrooms to see the number of occurrences of the words. They may also create their own corpus from their learners' writings or make their learners become aware of various text types and genres.

Furthermore, some of the platforms are interactive such as FLAX (Interactive Language Learning: FLAX library (nzdl.org), Wu, et. al., 2006) and Compleat Lexical Tutor (https://www.lextutor.ca/, Cobb, 2020). These platforms offer many activities as well as resources for English language learners at every level. Teachers can have access or create corpus thanks to these free concordancers while learners observe concordances, study patterns of language use in corpora and witness how a word may be used. Allowing learners to access corpora, DDL lets learners discover patterns and encounter various examples of authentic language uses of those patterns. In addition, learners become more aware of register and text types (e.g., speech or writing), frequency and typicality, and discourse and style. If learners are made aware of their own errors and underuse as well as overuses, their awareness can be raised, and they focus more on what they produce in their own productions.

As for the teachers who do not have any equipment and the Internet in their classrooms, some resources offer printed materials for DDL such as Tim Johns' Kibbitzers (https://lexically.net/TimJohns/), Lancaster University – Corpus for Schools: (Corpus for Schools | Corpus resources for A-level English Language and English Language Teaching (lancs.ac.uk), Gablasova, 2020), Creating Corpus-Informed Materials for the English as a Foreign Language Classroom

(https://pressbooks.pub/elenlefoll/, Le Foll, 2021). They also provide step-by-step guideline along with a lesson plan.

Sample worksheets along with a sample lesson plan about using language patterns (e.g., lexico-grammatical patterns, lexical phrases, making connections within and between sentences, expression of stance and voice, giving reasons, identifying causes and effects) and using corpora to write for different purposes have been adapted from Karpenko-Seccombe (2021) and can be found in Appendix 3. In the adaptations of the sample worksheets, the content and the format of the worksheets have been used while the tasks have been modified by revising as well as changing the corpus tools used.

Apart from DDL, there are also some corpus-informed textbooks which include real language uses. Some of these books are *Exploring Academic English* (Thurstun and Candlin, 1997), *Touchstone series* (McCarthy et al., 2005), *Real grammar* (Conrad & Biber, 2009), *Grammar and beyond* (Reppen et al.,2012). These can be especially useful addition to the usual textbooks, exposing learners to real language uses.

Last but not least, improvements in corpus tools in recent years does not appear to have encouraged ELT publishers or practitioners to put more emphasis on computer-retrieved formulaic language in the curriculum and/or materials. Materials and teachers lag behind the developments of the corpus studies, this is due to the gap between theory and practice and this gap may be shortened via making the teachers and instructors equip with corpus knowledge and practice. The way of doing this may be realized with required classroom materials. That's why, more publications with the integration of corpora applications into the teacher education and foreign language teaching should be made. Textbooks and other teaching materials should be prepared taking these frequently used structures into account. These language teaching materials can facilitate the acquisition of certain MWIs by several ways such as usage notes, drilling exercises, texts, or tapes with repeated instances of a word sequence, collocational observations and illustrations or just drawing a class's attention to a feature.

Another way of shortening the gap between corpus studies and teachers/instructors is the integration of corpus to the ELT departments for pre-service teachers. It may be added as a course to the existing curriculum. In addition, direct and indirect uses of corpora and data-driven learning and teaching in the classroom should be introduced to the in-service teachers through different in-service trainings.

6.5. Recommendation for further research

A great number of L2 studies on formulaic language have examined learner writing by comparing it to L1 expert writing, mainly research articles, and have focused on the features that deviate from L1 norms. There exists relatively little research using parallel corpora of native and nonnative academic genres, other than research articles. Future research targeting the same type of academic writing produced by different language groups would provide a more concrete picture of the groups' formulaic language use, as well as of the features specific to the register.

LBs comprise a large portion both in written and spoken discourse. They also have significant discourse functions. Mastering these expressions increase learners' proficiencies bringing them closer to produce more native like structures. Since the studies conducted so far have shown that different registers have different sets of lexical bundles, more lexical bundle studies should be conducted at different registers from different disciplines and genres. Besides, studies should be conducted more deeply in order to increase the reliability, objectivity, and validity. The quantitative analysis needs to be supported by qualitative analyzes considering examination of expanded concordance lines. More non-native corpora should be collected and studied to make better contrastive interlanguage analysis.

Also, the use of LBs by novice and expert writers from both Turkish writers and different L1 backgrounds should also be investigated whether to see the similarities and differences are language specific or at the interlanguage stages. How mother tongue and cultural writing styles influence the writers could also be revealed in this way. Also, much further study is needed especially from a psycholinguistic perspective. These expressions deserve more attention.

Additionally, studies with data-driven learning and teaching, focusing on these structures and expressions should be conducted at different disciplines and ELT departments. Studies aiming to increase awareness of L2 learners in the use of bundles and formulaic language are needed. Learners can benefit from the use of corpora and DDL activities through their education lives and can also pass those on to their own students in the future.

These types of studies to be conducted may help researchers understand the stages learner go through as well as their errors. Understanding how these structures are processed and where the problematic areas lie may provide opportunities to study more experimentally by helping learners acquire these structures and become more fluent and

proficient language learners.

6.6. Limitations

This dissertation has several limitations that should be considered in future research. First of all, it focused on Turkish EFL learners' LB productions in argumentative essays. Expressions are not that common and varied in learner writing. This may have had an influence on the results, distinguishing from other studies. Also, two of the corpora used are small sized corpora and small scale of corpora can present only limited findings. Therefore, the results of the present study cannot be generalized.

Secondly, even though erroneous uses are frequent in learner writing, we excluded them in this dissertation. That is why, learners' errors (i.e., omission, misuses) as well as their attempts to use different bundles have not been considered. Examining the erroneous structures contributes to the understanding learners' LB uses. The results would have differed had we not excluded the erroneous and repetitive structures from the analysis.

In addition, our study only included Turkish L1 learners and native speakers. The use of lexical bundles by L2 writers of different proficiency levels, the differences between expert and novice writing are also required to gain wider insights. Comparing the LBs used by learners with other language backgrounds would give more insights and helps us generalize language features.

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.

Aijmer, K. (2002). *English discourse particles: Evidence from a corpus*. John Benjamins Publishing.

Allen, D. (2009). Lexical bundles in learner writing: An analysis of formulaic language in the ALESS learner corpus. Komaba Journal of English Education, 1, 105-127.

Alonso, M. R. (2002). *The Role of Transfer in Second Language Acquisition*. *Servicio de Publicacións.* Vigo: Universidade.

Altenberg, B. (1993). Recurrent word combinations in spoken English. In: J. D. Arcy (Ed.), *Proceedings of the Fifth Nordic Association for English Studies Conference* (pp. 17–27). Reykjavik: University of Iceland.

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

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*, 173–195.

Anthony, L. (2006). Developing a freeware, multiplatform, corpus analysis tool kit for the technical writing classroom. *Transactions on Professional Communication, 49*(3), 25-32.

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

Aston, G. (2000). Corpora and language teaching. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 7–17). Frankfurt, Germany: Peter Lang.

Aston, G. (2001). *Learning with corpora*. Houston, TX: Athelstan.

Aybek S. , Can C. (2019). A Learner Corpus-Based Study on the Use of English Prepositional Verbs by Turkish EFL Learners. in: Contemporary Means and Methods in ELT and Applied Linguistics, Cem Can; Paschalia Patsala; Zoi Tatsioka, Editor, Lychnos-Printhouse, Tallinn, pp.51-77.

Barlow, M. (1996). Corpora for theory and practice. *International Journal of Corpus Linguistics, 1*, 1–37.

Barnbrook, G. (1996). *Language and Computers*. Edinburgh: Edinburgh University

Press.

Bernardini, S. (2000a). *Competence, capacity, corpora. A study in corpus-aided language learning*. Bologna, Italy: CLUEB.

Bernardini, S. (2000b). Systematising serendipity: Proposals for concordancing large corpora with language learners. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 225–234). Frankfurt, Germany: Peter Lang.

Bernardini, S. (2001) '"Spoilt for Choice": A Learner Explores General Language Corpora', in G. Aston (ed.) Learning with Corpora. Houston: Athelstan, pp. 220–49.

Bernardini, S. (2002). Exploring new directions for discovery learning. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 165–182). Amsterdam, The Netherlands: Rodopi.

Bernardini, S. (2004) 'Corpora in the Classroom: An Overview and Some Reflections on Future Developments', in J. Sinclair (ed.) *How to Use Corpora in Language Teaching. Amsterdam and Philadelphia*, PA: John Benjamins, pp. 15–36.

Bernardini, S. (2004a). Corpora in the classroom: An overview and some reflections on future developments. In J. M. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15–36). Amsterdam, The Netherlands: John Benjamins.

Bernardini, S. (2004b). Corpus-aided language pedagogy for translator education. In K. Malmkjaer (Ed.), *Translation in undergraduate degree programmes* (pp. 97–112). Amsterdam, The Netherlands: John Benjamins.

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics,* 14(3), 275-311.

Biber, D. and Reppen, R. (2002) 'What Does Frequency Have to Do with Grammar Teaching?' *Studies in Second Language Acquisition* 24: 199–208.

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

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

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

Biber, D., Conrad, S. Reppen, P. (1998). *Corpus linguistics: Investigating language structure and use.* Cambridge: Cambridge University Press.

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 (A. Wilson, P. Rayson, & T. McEnery, Eds.). Applied Linguistics, 25(3), 371–405.

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

Biber, D., Leech, G., & Conrad, S. (2002). *Longman student grammar of spoken and written English*. London, UK: Longman.

Bishop, H. (2004). *The noticing of formulaic sequences by second language readers*. Unpublished doctoral dissertation. USA: University of Wisconsin- Madison.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, H. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, 10, 245–261.

Boers, F., Eyckmans, J. Kappel, J., Stengers, H., & Demecheleer, M. (2016). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, 10(3), 245–261.

Bolander, M. (1989). Prefabs, patterns and rules in interaction? Formulaic speech in adult learners' L2 Swedish. In K. Hyltenstam & L. K. Obler (Eds.), *Bilingualism across the lifespan: Aspects of acquisition, maturity, and loss* (pp. 73–86). Cambridge: Cambridge University Press.

Boulton, A. (2008a) 'DDL: Reaching the Parts Other Teaching Can't Reach?' in A. Frankenburg-Garcia (ed.) *Proceedings of the 8th Teaching and Language Corpora Conference.* Lisbon: Associação de Estudos e de Investigação Cientifica do ISLA-Lisboa, pp. 38–44.

Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL, 21,* 37–54.

Boulton, A. (2009b) 'Data-driven Learning: Reasonable Fears and Rational Reassurance', **Indian Journal of Applied Linguistics** 35(1): 81–106.

Boulton, A. (2015). Applying data-driven learning to the web. In A. Le´nko-Szyma´nska

& A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 267–295). Amsterdam: John Benjamins.

Braun, S. (2005) .'*From Pedagogically Relevant Corpora to Authentic Language Learning Contents',* ReCALL 17(1): 47–64.

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 5.x. [software]. Available at: http://corpora.lancs.ac.uk/lancsbox.

Bychkovska, T., & Lee, J. (2017). At the same time: Lexical bundles in L1 and L2 university student argumentative writing. *Journal of English for Academic Purposes, 30*, 38-52.

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.

Can C., (2009). İkinci dil edinimi çalışmalarında bilgisiyar destekli bir Türk öğrenici İngilizcesi derlemi: ICLE'nin bir altderlemi olarak TICLE", *Dil Dergisi*, cilt.144, 16-34.

Can, C., (2016). *Learner Corpora and Data-driven Teaching*. Yeditepe University Graduate Seminars (pp.1). İstanbul, Turkey.

Can, C., (2017). A Learner Corpus-Based Study on Verb Errors of Turkish EFL Learners. *Journal of Education and Training Studies* , vol.5, 167-175.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English.* Cambridge, UK: Cambridge University Press.

Carter, R., Hughes, R., & McCarthy, M. (2000). *Exploring grammar in context. Grammar reference and practice.* Cambridge, UK: Cambridge University Press.

Chambers, A. (2005) 'Integrating Corpus Consultation in Language Studies', Language Learning and Technology 9(2): 111–25.

Chambers, A. (2019). Towards the corpus revolution? Bridging the research-practice gap. *Language Teaching*, 1-16. https://doi.org/10.1017/S0261444819000089

Chambers, A. and O'Sullivan, I. (2004) 'Corpus Consultation and Advanced Learners' Writing Skills', *ReCALL* 16(1): 158–72.

Chambers,A. (2007). Popularising corpus consultation by language learners and teachers. In E. Hidalgo, L. Quereda, & J. Santana (Eds.), *Corpora in the foreign language classroom* (pp. 3–16). Amsterdam, The Netherlands: Rodopi.

Chang, Y. & Swales, J. (1999). Informal elements in English academic writing: Threats

or opportunities for advanced non-native speakers? In K. Hyland & C. Candlin (Eds.), *Writing: Texts, processes and practices* (pp. 145-167). London: Longman.

Chao, P. (2010). *A study of collocation learning of junior high students in Taiwan via concordance.* Paper presented at the 2010 international conference on English teaching, Kaohsiung, Taiwan.

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

Chen, Y., & Baker, P. (2016). Investigating critical discourse features across second language development: Lexical bundles in rated learner essays, CEFR B1, B2 and C1. *Applied Linguistics, 37*(6), 849-880.

Cheng, W. (2007) 'Concgramming: A Corpus-driven Approach to Learning the Phraseology of Discipline specific Texts', *CORELL: Computer Resources for Language Learning* 1: 22–35.

Chenu, F., & Jisa, H. (2009). Reviewing some similarities and differences in L1 and L2 lexical development. A*cquisition et interaction en langue étrangère,* 17-38.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System, 25*, 301–315.

Cobb, T. (1999). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research & Development*, *47*, 15–33.

Cobb, T. (2020). Lextutor concordancer. Available online at lextutor.ca/conc/eng/

Colovic-Markovic, J. (2012). *The effects of explicit instruction of formulaic sequences on second-language writers.* Unpublished doctoral dissertation .Utah: The University of Utah.

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

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. (2007) . Teaching lexical bundles in the disciplines: An example from a writing intensive history class. *Linguistics and Education*, 17 (4) pp. 391–406.

Cortes, V., & Hardy, J.A. (2013). Analyzing the semantic prosody and semantic

preference of lexical bundles. In D. Belcher & G. Nelson (Eds.), *Critical and corpus-based approaches to intercultural rhetoric* (pp. 180-201). Ann Arbor, MI: University of Michigan Press.

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment.* Cambridge, U.K: Press Syndicate of the University of Cambridge.

Cowie, A. P. (1994). Phraseology. In R. E. Asher (Ed.), *The encyclopedia of language and linguistics* (pp. 3168–3171). Oxford: Pergamon.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213–238.

Coxhead, A., & Byrd, P. (2007). Preparing writing teachers to teach the vocabulary and grammar of academic prose. *Journal of Second Language Writing,* 16, 129–147.

Crossley, S., & Salsbury, T. (2011). The development of lexical bundle accuracy and production in English second language speakers. *IRAL: International Review of Applied Linguistics in Language Teaching,* 49, 1-26.

Crosthwaite, P. (2022). Data-driven learning for younger learners: Introducing CorpusMate. *Australian Linguistics Society (ALS) 2022*, Melbourne, Australia, 30 November-2 December 2022.

Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided error analysis. *Elsevier,* 26(2), 163-174.

Davies, Mark. (2008-) *The Corpus of Contemporary American English (COCA)*. Available online at https://www.english-corpora.org/coca/.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. Belgian Journal of English Language and Literatures (BELL), New Series (2), 225–246.

De Cock, S., Granger, S., Leech, G., & McEnery, T. (1998). An automated approach to the phrasicon of EFL learners. In S. Granger (Ed.), *Learner English on computer* (pp. 68–80). London & New York: Addison Wesley Longman.

Debrock, M., Flament-Boistrancourt, D. and Gevaert, R. (1999) 'Le manque de "naturel" des interactions verbales du non-francophone en français. Analyzes de quelques aspects à partir du corpus LANCOM', Faits de Langue 13 (Oral-Ecrit: Formes et théories): 46–56.

Dörnyei, Z., & Skehan, P. (2003). Individual differences in second language learning. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 589–630). Malden, MA: Blackwell.

Ellis, N. (2002) 'Frequency Effects in Language Processing. A Review with Implications

for Theories of Implicit and Explicit Language Acquisition', *Studies in Second Language Acquisition* 24: 143–88.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition, 18*, 91–126.

Ellis, R. (2005). Measuring Implicit and Explicit Knowledge of a Second Language: A Psychometric Study. *Studies in Second Language Acquisition*, 27(2), 141-172.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second-language speakers: Psycholinguistics, corpus Linguistics, and TESOL. *TESOL Quarterly*, 42, 375–396.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, 32, 17–44.

Elturki, E. (2015). *The Development of Formulaic Sequences: A Longitudinal Learner Corpus Investigation.* Dissertation, Washington State University, Washington.

Ergin, I. (2013). *The effect of explicit teaching of formulaic language on Turkish English as a foreign language (EFL) learners' writing performance*. Unpublished master's thesis. Ankara: Bilkent University.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. **Text,** 20(1), 29–62.

Farr, F. (2008) 'Evaluating the Use of Corpus-based Instruction in a Language Teacher Education Context: Perspectives from the Users', *Language Awareness* 17(1): 25–43.

Firth, J. R. (1951). Modes of meaning. In J. R. Firth (Ed.), *Essays and studies* (pp. 118–149). London: Oxford University Press.

Firth, J. R. (Ed.), (1957). *Papers in linguistics 1934–1951*. Oxford: Oxford University Press.

Fligelstone, S. (1993) 'Some Reflections on the Question of Teaching, from a Corpus Linguistics Perspective', *ICAME Journal* 17: 97–110.

Flowerdew, L. (2001). The exploitation of small learner corpora in EAP materials design. In M. Ghadessy, A. Henry, & R. L. Roseberry (Eds.), *Small corpus studies and ELT. Theory and practice* (pp. 363–380). Amsterdam, The Netherlands: John Benjamins.

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

Gabrielatos, C. (2005) ‘Corpora and Language Teaching: Just a Fling or Wedding Bells?’ *TESL-EJ* 8(4): 1–35

Gavioli, L. (2006). *Exploring corpora for ESP learning*. Amsterdam, The Netherlands: John Benjamins.

Gavioli, L. and Aston, G. (2001) ‘Enriching Reality: Language Corpora in Language Pedagogy’, *ELT Journal 5*5(3): 238–46.

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

Gilquin, G. (2007). To err is not all. What corpus and elicitation can reveal about the use of collocations by learners. *Zeitschrift für Anglistik und Amerikanistik, 55*(3), 273-291.

Goodale, M. (1995). *Collins COBUILD concordance samplers 4: Tenses*. London: Harper-Collins.

Granath, S. (2009). Who benefits from learning how to use corpora? In Aijmer, K. (Ed.), *Corpora and Language Teaching* (pp.47-66). Amsterdam: John Benjamins Publishing Company.

Granger, S. & S. Tyson (1996). Connector Usage in Native and Non-native English Essay Writing. *World Englishes 15(1),* 17-27.

Granger, S. (1996). From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora. Languages in Contrast: Text-based cross-linguistic studies, Lund Studies in English 88, *Lund University Press*, 37– 51.

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: Clarendon Press.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (ed.) *Learner English on Computer. Addison Wesley* Longman : London & New York, 3-18.

Granger, S. (2002). A Bird’s-eye view of learner corpus research. in computer learner corpora, econd language acquisition and foreign language teaching. Granger, S. & Hung, J. & Petch-Tyson, S. (eds), 3-33. John Benjamins, Amsterdam.

Granger, S. (2003) ‘The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research’, *TESOL Quarterly*, 37(3): 538–46.

Granger, S. (2009). Commentrary on part I: Learner corpora: A window onto the L2

phrasicon. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 61-65). New York, NY: Palgrave Macmillan.

Granger, S., & Paquot, M. (2008). Disentangling the phraseological web. In S. Granger & Meunier (Eds.), *Phraseology: An interdisciplinary perspective*. Amsterdam: John Benjamins.

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

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. In S. Granger and F. Meunier (Eds.), *Phraseology: An interdisciplinary perspective* (pp. 3-25). John Benjamins Publishing Company.

Harwood, N. (2005). 'Nowhere has anyone attempted…In this article I aim to do just that': A corpus-based study of self-promotional *I* and *we* in academic writing across four disciplines. *Journal of Pragmatics, 37*(8), 1207-1231.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: a study into the ways Norwegian students cope with English vocabulary. *International Journal of Applied Linguistics,* Volume 4, Issue 2, 237-258.

Haswell, R. (1991). *Gaining ground in college writing: Tales of development and interpretation.* Dallas: Southern Methodist University Press.

Hinkel, E. (2002). Second language writers' text. Mahwah, NJ: Erlbaum.

Hoey, M. (1997). *From concordance to text structure: New uses for computer corpora. In Practical Applications in Language Corpora.* (eds), 2–23. Łódź: Łódź University Press.

Hoey, M. (2004). "Lexical priming and the properties of text". In Aston, G., Bernardini, S., Steward, D. (2004) (eds.). *Corpora and language learners. Language, 83*(4).

Hoey, M. P. (2005). *Lexical priming. A new theory of words and language.* London, UK: Routledge.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S., & Francis, G. (2000). *Pattern grammar. A corpus-driven approach to the lexical grammar of English*. Amsterdam, The Netherlands: John Benjamins.

Hunston, S., Francis, G., Manning, E. (1997). Grammar and vocabulary: Showing the connections. *ELT Journal 5*1 (3).

Hyland, K. (2004). *Disciplinary discourses: social interactions in academic writing*. USA: University of Michigan Press.

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(1), 41–62.

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

Jalali, H., Rasekh, A.E., & Rizi, M.T. (2008). Lexical bundles and intradisciplinary variation: The case of applied linguistics. *International Journal of Language Studies,* 2 (4), 447-484.

Jaworska, S., Krummes, C., & Ensslin, A. (2015). Formulaic sequences in native and non-native argumentative writing in German. *International Journal of Corpus Linguistics, 20*(4), 500-525.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *The Modern Language Journal, 91*, 433–445.

Johns, T. (1991a) 'Should You Be Persuaded', *ELR Journal* 4: 1–16.

Johns, T. (1991b) 'From Printout to Handout: Grammar and Vocabulary Teaching in the Context of Data-driven Learning', *English Language Research Journal 4*: 27–45.

Johns, T. F. (1986). Microconcord: A language-learner's research tool. *System, 14*, 151–162.

Johns, T. F. (1994). From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 27–45). Cambridge, UK: Cambridge University Press.

Johns, T. F. (1997). Contexts: The background, development, and trialing of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). London, UK: Longman.

Johns, T. F. (2002). Data-driven learning: The perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 107–117). Amsterdam, The Netherlands: Rodopi.

Johns, T. F., Hsingchin, L., & Lixun, W. (2008). Integrating corpus-based CALL programs in teaching English through children's literature. *Computer Assisted Language Learning, 21,* 483–506.

Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences. N. Schmitt (Ed.), in *Formulaic sequences acquisition, processing and use*, (pp.269-292). Philadelphia: John Benjamins.

Jones, R. L. (1997). Creating and using a corpus of spoken German. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 146–156). London, UK: Longman.

Juknevičienė, R. (2009). Lexical bundles in learner language: Lithuanian learners vs. Native speakers. *KaLBOTYRa,* 61(3), 61-72.

Juknevičienė, R. (2011). *Lexical bundles in non-native speaker and native speaker written English.* Unpublished doctoral dissertation. Vilnius University.

Kaltenbock, G. and Mehlmauer-Larcher, B. (2005) 'Computer Corpora and the Language Classroom: The Language Classroom: On the Potential and Limitations of Computer Corpora in Language Teaching', *ReCALL* 17(1): 65–84.

Karabacak, E. & Qin, J. (2013). Comparison of lexical bundles used by Turkish, Chinese, and American university students. *Procedia-Social and Behavioral Sciences*, 70, 622-628.

Karpenko-Seccombe, T. (2021). *Academic Writing with Corpora: A Resource Book for Data-Driven Learning.* London and New York: Routledge, Taylor and Francis Group.

Kazemi, M., Katiraei, S., & Rasekh, A. E. (2014). The Impact of Teaching Lexical Bundles on Improving Iranian EFL Students' Writing Skill. *Procedia-Social and Behavioral Sciences*, 98, 870-875.

Kennedy, G. (1992). Preferred ways of putting things with implications for language teaching. In J. Svartvik (Ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4–8 August 1991* (pp. 335–373). Berlin, Germany: Mouton de Gruyter.

Kennedy, G. (2003) 'Amplifier Collocations in the British National Corpus: Implications for English Language Teaching', *TESOL Quarterly* 37(3): 467–87.

Keshavarz, M. H. (2012). Sources of Errors. In *Contrastive Analysis and Error Analysis* (pp. 117-128). Rahnama Press.

Kettemann, B. (1995). On the use of concordancing in ELT. *Arbeiten aus Anglistik und Amerikanistik, 20*, 29–41.

Kilimci, A., & Can, C., (2008). TICLE: Uluslararası Türk Öğrenici İngilizcesi Derlemi. *22nd Ulusal Dilbilim Kurultayı* (pp.1-11). Van, Turkey.

Krashen, S. (1983). *The Input Hypothesis.* London: Longman.

Krashen, S. (1988). *Second Language Acquisition and Second Language Learning.* London: Prentice-Hall.

Latifi, M. A., & Afraz, S. (2015). The Effect of Explicit Instruction of Lexical Bundles on Developing Writing Skill in Pre-Intermediate EFL Learners. *International Journal of Review in Life Sciences*., 5(8), 84-94.

Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography*, 24, 29–49.

Laufer, B., & Waldman, T. (2011). Verb- noun collocations in second language writing: A corpus analysis of learners' *English. Language Learning,* 61(2), 647-672.

Le Foll, E. (2021). *Creating Corpus-Informed Materials for the English as a Foreign Language Classroom.* Pressbooks.

Leech, G. (1997) 'Teaching and Language Corpora: A Convergence', in A. Wichmann, S. Fligelstone, A. M. McEnery and G. Knowles (eds) *Teaching and Language Corpora. London: Addison Wesley* Longman, pp. 1–23.

Li, J. & Schmitt, N. (2009). The acquisition of lexical phrases in academic writing: A longitudinal case study. *Journal of Second Language Writing*, 18 (2), 85-102.

Liu, D. (2012). The most frequent multiword constructions in academic written English; A multi-corpus study. *English for Specific Purposes, 31*(1), 25–35.

Long, M. H. (1988). Instructed Interlanguage Development. In L. Beebe (Ed.), Issues in Second Language Acquisition: *Multiple Perspectives* (pp. 115-141). Rowley, MA: Newbury House.

Lorenz, G. (1999). *Adjective Intensification – Learners vs Native Speakers. A Corpus Study of Argumentative Writing.* Amsterdam & Atlanta: Rodopi.

Lüdeling, A., & Kytö, M. (2009). *Corpus Linguistics An International Handbook.* Handbooks of Linguistics and Communication Science.

Mair, C. (2002) 'Empowering Non-native speakers: The Hidden Surplus Value of Corpora in Continental English Departments', in B. Kettemann and G. Marko (eds) *Teaching and Learning by Doing Corpus Analysis*. Amsterdam: Rodopi, pp. 119–30.

Martinez, R., & Schmitt, N. (2012). Phrasal expressions list. *Applied Linguistics,* 33(3), 299-320.

McCarthy, M. and Carter, R. (2004) 'Size Isn't Everything: Spoken English, Corpus, and

the Classroom', *TESOL Quarterly* 35 (2): 337–40.

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone student's book 1.* Cambridge, UK: Cambridge University Press.

McEnery, T. & Wilson, A. (2001). *Corpus Linguistics (Second Ed.).* Edinburg University Press.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies.* London: Routledge.

Meunier, F. (2002). The Pedagogical Value of Native and Learner Corpora in EFL Grammar Teaching. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, pp. 364-380). London & New York: Routledge.

Meunier, F. (2002). The Pedagogical Value of Native and Learner Corpora in EFL Grammar Teaching. In E. Hinkel (Ed.), *Handbook of Research in Second Language Teaching and Learning* (Vol. 2, pp. 364-380). London & New York: Routledge.

Meunier, F., & Granger, S. (Eds.). (2008). *Phraseology in foreign language learning and teaching*. Amsterdam, The Netherlands: John Benjamins.

Milton, J. (1997) 'Providing Computerized Self-access Opportunities for the Development of Writing Skills', in P. Benson and P. Voller (eds) Autonomy and Independence in Language Learning. *Harlow: Longman,* pp. 204–14.

Mindt, D. (1981). Angewandte Linguistik und Grammatik f¨ur den Englischunterricht. In P. Kunsmann & O. Kuhn (Eds.), *Weltsprache Englisch in Forschung und Lehre: Festschrift f¨ur Kurt W¨achtler* (pp. 175–186). Berlin, Germany: Schmidt.

Mindt, D. (1987). *Sprache—Grammatik—Unterrichtsgrammatik. Futurischer Zeitbezug im Englischen I*. Frankfurt, Germany: Diesterweg.

Mindt, D. (1997). Corpora and the teaching of English in Germany. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 40–50). London, UK: Longman.

Moon, R., (1997). *Vocabulary connections: Multi-word items in English. In: Norbert Schmitt, Michael McCarthy: Vocabulary: Description, Acquisition and Pedagogy.* Cambridge University Press, Cambridge 1997, 40–63.

Morris, L. and Cobb, T. (2004). Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System 32*: 75–87.

Mukherjee, J. (2006) 'Corpus Linguistics and Language Pedagogy: The State of the Art – and Beyond', in S. Braun, K. Kohn and J. Mukherjee (eds) *Corpus Technology and Language Pedagogy. Frankfurt am Main: Peter Lang*, pp. 5–24.

Mukherjee, J., & Rohrbach, J. (2006). Rethinking Applied Corpus Linguistics from a Language-pedagogical Perspective: New Departures in Learner Corpus Research. In: Kettemann, Bernhard/Marko, Georg (eds.). *Planing, Gluing and Painting Corpora*, 205-232.

Muşlu, M. (2015). *Stance Lexical Bundles in Academic L2 Writing: A Contrastive Interlanguage Analysis.* PhD Dissertation, Cukurova University, Adana.

Nation, P. (2001). *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Neely, E. & Cortes, V. (2009). A little bit about: analyzing and teaching lexical bundles in academic lectures. *Language Value*, 1(1), 17-38.

Neff, J. (2008). Contrasting English-Spanish interpersonal discourse phrases. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 85-100). Amsterdam: John Benjamins.

Nekrasova, T.M. (2009). English L1 and L2 Speakers' Knowledge of Lexical Bundles. *Language Learning*, 59 (3), 647-686.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics*, 11 (3), 283-304.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied linguistics*, 24(2), 223-242.

Nesselhauf, N. (2004) 'Learner Corpora and their Potential for Language Teaching', in J. Sinclair (ed.) How to Use Corpora in Language Teaching. Amsterdam: John Benjamins, pp. 125–52.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam, The Netherlands: John Benjamins.

Nguyen, H. T. H. (2014). *The acquisition of formulaic sequences in high-intermediate ESL learners*. Unpublished doctoral dissertation. Pennsylvania: University of Pennsylvania.

Nicholls, D. (2003). The Cambridge Learner Corpus: Error Coding and Analysis for Lexicography and ELT. In: Archer, D./Rayson, P./Wilson, A./McEnery, T. (eds.). *Proceedings of the Corpus Linguistics 2003 Conference.* UCREL, Lancaster University, 572-581.

O'Keeffe, A., McCarthy, M. J. and Carter, R. A. (2007). *From Corpus to Classroom.* Cambridge: Cambridge University Press.

O'Sullivan, Í. (2007) 'Enhancing a Process-Oriented Approach to Literacy and Language Learning: The Role of Corpus Consultation Literacy', *ReCALL 1*9(3): 269–86.

Osborne, J. (2004) 'Top-down and Bottom-up Approaches to Corpora in Language Teaching', in U. Connor and T. A. Upton (eds) Applied Corpus Linguistics. *A Multidimensional Perspective. London: Rodopi,* pp. 251–65.

Öztürk, Y. (2014). *Lexical bundle use of Turkish and native English writers: A corpus-based study.* Unpublished MA thesis. Eskişehir: Anadolu University Graduate School of Educational Sciences.

P´erez-Paredes, P. (2003). Small corpora as assisting tools in the teaching of English news language: A preliminary tokens-based examination of Michael Swan's *Practical English Usage* news language wordlist. *ESP World.*

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics,* 32, 130–149.

Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). New York, NY: Longman.

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.

Peters, E. (2012). Learning German formulaic sequences: The effect of two attention drawing techniques. *Language Learning Journal,* 40, 65–79.

Peters, E. & Pauwels, P. (2015). Learning academic formulaic sequences. *Journal of English for Academic Purposes,* 20, 28-39.

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

Quinn, C. (2015). Training L2 writers to reference corpora as a self-correction tool. *ELT Journal, 69*(2), 165–177.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*. 13:4 pp. 519-549.

Rayson, P. and Garside, R. (2000). Comparing corpora using frequency profiling. In proceedings of the workshop on Comparing Corpora, held in conjunction with the

38th annual meeting of the Association for Computational Linguistics (ACL 2000). 1-8 October 2000, Hong Kong, pp. 1 - 6.

Renouf, A. & J. M. Sinclair (1991). Collocational frameworks in English. In Aijmer, K. & B. Altenberg (eds.) *English Corpus Linguistics,* 128–143. London: Longman.

Renouf, A. (1997) 'Teaching Corpus Linguistics to Teachers of English', in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds) *Teaching and Language Corpora.* London: Longman, pp. 255–66.

Reppen, R. (2009). Exploring L1 and L2 writing development through collocations: A corpus-based look. In A. Barfield & H. Gyllstad (Eds.), *Researching collocations in another language: Multiple interpretations* (pp. 49-59). New York, NY: Palgrave Macmillan.

Robinson, P., & Ellis, N. C. (2008). Conclusion: Cognitive linguistics, second language acquisition and L2 instruction – issues for research. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 489-545). New York, NY: Routledge.

Römer, U. (2004a). Comparing real and ideal language learner input. The use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 151–168). Amsterdam, The Netherlands: John Benjamins.

Römer, U. (2005) *Progressives, Patterns, Pedagogy: A Corpus-driven Approach to English Progressive Forms, Functions, Contexts, and Didactics.* Amsterdam: John Benjamins.

Römer, U. (2006) 'Pedagogical Applications of Corpora: Some Reflections on the Current Scope and a Wish List for Future Developments', *Zeitschrift für Anglistik und Amerikanistik* 54(2): 121–34.

Römer, U. (2007). Looking at *looking:* Functions and contexts of progressives in spoken English and "school" English. In W. Teubert & R. Krishnamurthy (Eds.), *Corpus linguistics. Critical concepts in linguistics* (Vol. *4*, pp. 3–14). London, UK: Routledge. (Reprinted from *The changing face of corpus linguistics* by A. Renouf & A. Kehoe, Eds., 2006, Amsterdam, The Netherlands: Rodopi)

Römer, U. (2008). Corpus research and practice:What help do teachers need and what can we offer? In: Aijmer, Karin (ed.), *Teaching and Corpora.* Amsterdam: John Benjamins.

Römer, U. (2011). *Corpus Research Applications in Second Language Teaching Annual Review of Applied Linguistics* 31, Cambridge University Press, 205–225.

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

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

Schmitt, N. & Grandage, S. & Adolphs, S. (2004). Are corpus-derived recurrent clusters psycholinguistically valid? In N. Schmitt (Ed.), Formulaic sequences: Acquisition, processing, and use (pp. 127-152). Amsterdam: Benjamins.

Schmitt, N., (2000). *Vocabulary in language teaching*. Cambridge, UK: Cambridge University Press.

Schmitt, N., Dörnyei, Z., Adolphs, S., & Durow, V. (2004). Knowledge and acquisition of formulaic sequences: A longitudinal study. In N. Schmitt (Ed.), T*he acquisition and use of formulaic sequences* (pp. 55-86). Amsterdam: John Benjamins Publishing Company.

Scott, M. & Thompson, G. (eds.) (2001). *Patterns of Text: In Honour of Michael Hoey*. Amsterdam: John Benjamins.

Scott, M., (2022), *WordSmith Tools version 8* (64-bit version) Stroud: Lexical Analysis Software.

Scott, M., & Tribble, C. (2006). *Textual Patterns: Key Words and Corpus Analysis. In Language Education* (Vol. 11, No. 22). Philadelphia: John Benjamins.

Selinker, L. (1972). Interlanguage. *IRAL - International Review of Applied Linguistics in Language Teaching*, *10*(1-4).

Shin, Yu Kyoung, "*Lexical Bundles in Argumentative Essays by Native and Nonnative English-Speaking Novice Academic Writers."* Dissertation, Georgia State University, 2018.

Simpson-Vlach, R. C., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*, 487–512.

Sinclair, J. (1998), 'The lexical item', in E. Weigand (ed.), *Contrastive Lexical Semantics*. Amsterdam and Philadelphia: *John Benjamins,* pp. 1–24.

Sinclair, J. (2004). Trust the Text: Language, Corpus and Discourse. London: Routledge.

Sinclair, J. M. (2004c). New evidence, new priorities, new attitudes. In J. M. Sinclair

(Ed.). *How to use corpora in language teaching* (pp. 271–299). Amsterdam, The Netherlands: John Benjamins.

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

Sinclair, J. M. (Ed.). (1990). *Collins COBUILD English grammar*. London: HarperCollins.

Sinclair, J. M. (Ed.). (1992). *Collins COBUILD English usage*. London: HarperCollins.

Sinclair, J. M. (Ed.). (2001). *Collins COBUILD English dictionary for advanced learners*. London: HarperCollins.

Sinclair, J. McH. (2003) *Reading Concordances.* London: Longman.

Sinclair, J.M. (1991). *Corpus concordance collocation.* Oxford, UK: Oxford University Press.

St John, E. (2001) 'A Case for Using a Parallel Corpus and Concordance for Beginners of a Foreign Language', *Language Learning and Technology* 5(3): 185–203.

Staehr, L. S. (2009). Vocabulary knowledge and advanced listening comprehension in English as a foreign language. *Studies in Second Language Acquisition*, 31, 577–607.

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

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford: Blackwell.

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics* 7(2): 215–44.

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.

Sun, Y.-C. (2003) 'Learning Process, Strategies and Web-based Concordancers: A Case Study', *British Journal of Educational Technology* 34(5): 601–13.

Sun, Y-C., & Wang, L-Y. (2003). Concordancers in the EFL classroom: Cognitive

approaches and collocation difficulty. *Computer Assisted Language Learning, 16*(1), 83-94.

Svensen, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making.* Cambridge: Cambridge University Press.

Taylor, J. (2012). *The mental corpus: How language is represented in the mind*. Oxford, UK: Oxford University Press.

Thomas, D. (2005). *Type Token Ratios in One Teacher's Classroom Talk: An Investigation of Lexical Complexity.* Retrieved 16 December 2014,

Thornbury, S. (2002). *How to teach vocabulary.* London: Longman.

Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work.* Amsterdam and Philadelphia: Benjamins

Tolchinsky, L. (2006). The emergence of writing. In C. A. McArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of writing research*. (pp. 83-95). New York: The Guilford.

Tono, Y. (2011). TALC in action: Recent innovations in corpus-based English language teaching in Japan. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 3–25). London, UK: Continuum.

Uçar, S. (2007). A Corpus-Based Study on the Use of Three-Word Lexical Bundles in the Academic Writing by Native English and Turkish Non-Native Writers. *English Language Teaching*, Vol. 10, N. 12, p. (28-36).

Varantola, K. (2000) 'Translators, Dictionaries and Text Corpora', in S. Bernardini and F. Zanettin (eds) *I Corpora nella Didattica della Traduzione.* Bologna: CLUEB, pp. 117–33.

Vidaković, I. & Barker, F. (2009). Lexical development across second language proficiency levels: a corpus-informed study. In R. M. Millar & M. Durham (Eds.), *Language, Learning & Context: Proceedings of the 42nd Annual Meeting of the British Association for Applied Linguistic*s (pp. 143-146). London: Scitsiugnil Press.

Walsh, S. (2007) 'Learner Dissertation Writing Corpus', unpublished.

Wang, Y., & Shaw, P. (2008). Transfer and universality: Collocation use in advanced Chinese and Swedish learner English. *ICAME Journal,* 32, 201–232.

Warren, M. J., & Greaves, C. (2007). *A corpus-driven approach to learning and teaching the communicative role of discourse intonation*. Paper presented at International

Symposium on Teaching English at Tertiary Level.

Webb, S., Newton, J., & Chang, A. C. S. (2013). Incidental learning of collocation. *Language Learning*, 63(1), 91–120.

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

Williams, G. (2003). From meaning to words and back: Corpus linguistics and specialized lexicography. Asp, 39-40, 10 February 2011.

Williamson, G. (2009). http://www.sltinfo.com/wp-content/uploads/2014/01/typetoken-ratio.pdf. Retrieved 3 August 2013.

Williamson, G. (2013). http://www.sltinfo.com/type-token-ratio.html Retrieved 10 July 2012.

Willis, D. (1990). *The lexical syllabus. A new approach to language teaching*. London, UK: HarperCollins.

Willis, D. and Willis, J. (2007) *Doing Task-based Teaching.* Oxford: Oxford University Press.

Willis, D., & Willis, J. (1989). *Collins COBUILD English course*. London, UK: HarperCollins.

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

Wood, D. & Namba, K. (2013). Focused instruction of formulaic language: Use and awareness in a Japanese university class. *The Asian Conference on Language Learning Official Conference Proceedings 2013*, pp. 203–212.

Wood, D. C. & Appel, R. (2014). Multiword constructions in first year university textbooks and in EAP textbooks. *Journal of English for Academic Purposes, 15,* 1–13.

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

Wray, A. & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language and Communication, 20,* 1–28.

Wray, A. (2000). Formulaic sequences in second language teaching: Principles and practice. *Applied Linguistics*, 21(4), 463-489.

Wu, S., Witten, I. H., & Franken, M. (2010). Utilizing lexical data from a web-derived corpus to expand productive collocation knowledge. *ReCALL*, 22, 83–102.

Wu, S., Witten, I., & Witten, I. (2006). *Towards a Digital Library for Language Learning.*

In Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries. Villach, Austria.

Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology, 12*, 31–48.

Yoon, H. and Hirvela, A. (2004) 'ESL Student Attitudes toward Corpus Use in L2 Writing', *Journal of Second Language Writing* 13: 257–83.

Yorio, C. (1980). Conventionalized language forms and the development of communicative competence. *TESOL Quarterly, 16*(4), 433–442.

APPENDICES

APPENDIX A. LEXICAL BUNDLES IN TICLE

Freq.	Range	N-grams
126	43	men and women
123	71	a lot of
120	79	it is not
119	75	in order to
119	94	on the other hand
119	63	they do not
109	76	the most important
108	69	in the world
106	76	there is no
98	64	most of the
86	62	there is a
83	33	the real world
79	62	as a result
76	60	it is a
72	47	there are some
71	54	do not have
70	49	in my opinion
69	49	it is the
67	43	the people who
65	61	first of all
65	40	people who are
65	49	they are not
64	50	that it is
62	43	there are many
59	46	i want to
58	48	because of the
57	40	in our country
53	37	can not be
53	35	they can not
51	38	to have a
51	26	women and men
50	40	day by day
33	37	should not be+
49	31	in the society
49	44	that they are
47	37	is not a
45	38	is the most

44	38	but it is
44	31	prepare students for
43	28	i believe that
43	35	that there is
43	31	the right to
43	27	university degrees are
41	34	according to the
41	35	they want to
41	36	in addition to
41	27	in the past
41	25	the real life
40	35	is one of the
40	32	the help of
40	40	to sum up
39	28	i do not
39	32	to be a
38	26	if there is
37	34	point of view
36	29	in the future
36	31	they have to
35	30	look at the
35	28	some of them
34	27	this is a
33	31	and they are
33	22	if you have
51	22	it should be
33	23	there are a lot
32	28	all over the world
32	26	as a result of
32	26	it will be
32	27	part of the
32	25	we can not
31	26	and do not
31	28	because they are
31	26	have the same
31	26	it is very
31	26	you can not
31	27	the fact that
31	28	there will be
30	27	we can see
30	27	one of them
30	25	that there are
30	20	this kind of
30	26	we look at

45	20	all of the+
16	20	all of them
29	23	you do not
29	22	do not want
29	27	i think that
29	19	in our society
29	23	you want to
29	27	most of them
29	24	the importance of
29	29	this is not
29	23	with each other
27	25	and it is
27	25	as it is
27	25	because of this
27	23	for example in
27	20	in the family
27	23	know how to
27	25	that they can
27	22	the number of
27	16	the other people
27	19	the students to
27	23	they think that
27	19	we do not
26	23	a person who
26	21	if you are
26	24	it can be+V3
26	25	in front of
26	22	there is not
26	26	this is the
26	20	to be able to
25	20	a chance to
25	23	important thing is
25	21	they have a
24	20	and they can
24	20	at the end
24	22	at the same time
24	19	both men and women
24	19	they should be
24	13	his or her
24	21	is not enough
24	19	money is the
24	20	there should be
24	19	to go to
23	8	animal testing is

23	19	aware of the
23	23	for example the
23	18	for the people
23	20	for the students
23	15	have the right
23	21	some of the
23	20	we can say
23	19	we see that
22	22	as a conclusion
22	20	because it is
22	18	more and more
22	19	people who have
22	18	that women are
22	21	the beginning of
22	17	the use of
22	15	the value of
22	21	there is an
22	17	they start to
22	17	to find a
22	16	to live in
22	17	we have to
25	17	they will not
21	18	and the other
21	16	at this point
21	15	every part of
21	18	for this reason
21	18	get rid of
21	16	how to use
21	16	in other words
21	19	is not the
21	19	is the most important
21	19	they can be
21	19	if they are
21	16	most of the people
21	19	people do not
21	19	that they have
21	12	the people who are
21	19	they can do
21	20	they try to
21	18	they will be
21	19	to make the
21	19	when they are
21	18	when we look at
20	15	a kind of

20	17	all the people
20	17	and so on
20	19	as much as
20	18	as well as
20	19	as you see
39	19	but there is
20	17	have to be
20	19	i think the
20	17	in spite of
20	18	related to the
20	14	the result of
20	16	the root of
20	15	there is nothing
20	17	to each other
20	18	to say that
20	16	will be a
19	16	it has been
19	16	it is also
19	17	it means that
19	11	sex equality is
19	14	thought to be
18	15	all around the world
18	15	in terms of
18	8	in the exams
18	12	in the house
18	16	in this way
18	13	it is so
18	11	you do not
18	7	you have money
17	12	in addition to this
17	13	can be done
17	14	do you think
17	15	i can say
17	15	person who is
17	14	should be given+
17	14	if they do
17	14	the one who
17	16	the other one
17	15	the rest of
17	15	should not be+adj
17	15	there is no need
17	17	what to do
17	15	when it is
16	15	students do not

16	15	that is the
16	15	all of them
16	13	for a long time
16	12	people think that
16	10	for real world
16	12	universities do not
16	11	all of them
16	12	many people who
16	13	for the sake of
16	10	in real life
16	8	in the real world
16	15	in this situation
16	13	we know that
15	16	men and women have
15	12	we can say that
15	14	although it is
15	13	if it is
15	13	but it is not
15	11	euthanasia is the
15	11	they do not
15	9	in the universities
15	12	we are not
15	14	the important thing is
15	15	more important than
15	13	there have been
14	11	in our lives
14	14	if we think
14	13	according to me
14	14	it does not
14	13	it should not be
14	12	some people who
14	11	there are lots of
14	12	we do not have
14	12	women can not
14	12	if they have
14	10	women can not
14	13	women who are
14	10	you can see
13	10	it can be+adj
13	10	it is true
13	12	as I said
13	12	as I mentioned
13	12	this can be
13	12	as we know

13	11	at the beginning
13	12	at the end of
13	12	in the end
13	13	if there is a
13	13	one of them is
13	13	they dont have
13	13	since they are
13	13	can be used
13	13	so there is
13	12	so they can
13	12	so it is
13	13	this means that
12	13	we all know
12	13	in order not to
12	13	you have a
12	12	in addition to these
12	12	I think it is
12	11	in fact it is
12	9	in the field
12	13	I will be
12	12	ı can say that
12	12	in the life
12	10	in this century
12	10	that it is a
12	11	that we have
12	11	you can do
12	11	that you are
12	11	it is possible
12	11	it is impossible
12	11	if they dont
12	10	there are many people
12	10	I have to
12	10	they need to
12	9	to be killed
12	11	we all know
12	11	if we look
12	11	whatever they want
11	11	but they are not
11	11	it is an
11	11	they should not
11	11	however it is
11	9	in a society
11	11	in front of the
11	11	it is not a

11	8	it is necessary
11	8	they do not know
11	8	we can give
11	10	it is not enough
11	8	students who are
11	5	the teachers should
11	8	there are also
11	10	there are so many
11	9	they do not want
11	9	he can not
11	9	they should not be
11	9	they have got
11	11	they should not
11	10	when they want
10	12	it is a fact
10	9	when we think
10	9	be used in
10	8	euthanasia should be
10	7	he wants to
10	9	if there are
10	9	if there is no
10	9	it is better
10	7	it is not a
10	10	it is obvious that
10	7	it is true that
10	8	it is used
10	10	people are not
10	10	they had to
10	8	they may have
10	7	to be used
10	7	what they have
10	7	a lot of people
10	7	they will have
10	7	they have some
10	7	you think that
10	7	we think that
10	7	women should be
10	7	I cannot

APPENDIX B. LEXICAL BUNDLES IN LOCNESS

Freq.	Range	N-grams
71	50	the death penalty
71	7	the fact that
63	44	one of the
53	52	in the united states
51	30	the right to
50	30	in order to
48	53	it is a/an
48	30	the united states
47	35	because of the
47	32	that it is
45	32	the use of
40	27	there is a/an
40	46	as well as
39	26	it is not
36	43	in public schools
35	26	to be able to
35	32	there is no
33	32	that they are
31	27	this is a
31	6	part of the
30	31	should not be+V3
29	25	a lot of
29	26	all of the
29	22	invention of the
29	29	men and women
28	25	due to the
28	24	the idea of
28	22	to be a
27	22	they do not
26	14	he or she
26	17	out of the
26	23	the number of
24	22	his or her
24	23	there are many
23	27	have to be
23	5	in favor of
23	11	in the past
23	24	is not the
23	11	the amount of
23	15	this type of

23	21	to be the
22	17	as long as
22	23	is not a
22	14	it would be
22	17	on the other hand
22	16	the case of
22	11	the end of
22	20	to have a
21	19	is a very
21	16	it is the
20	9	you do not
20	20	a part of
20	20	according to the
20	17	i feel that
20	19	in the case
20	20	they want to
19	18	seem(s) to be
19	20	because it is
19	18	because they are
19	14	in the world
19	20	most of the
19	15	some of the
19	18	the root of
19	14	way of life
19	16	whether or not
19	16	would not be
18	15	the right/s of
18	16	one of the most
18	17	the effects of
18	17	the majority of
18	14	the most important
17	17	many of the
17	15	the death penalty is
17	17	to make a
16	17	is one of the
16	16	is that the
16	18	it does not
16	13	it has been
16	8	such as the
16	9	the lack of
16	14	they are not
16	14	this is not
16	15	this is the
16	14	to say that

15	16	as a result of
15	12	be allowed to
15	14	do not have
15	16	look at the
15	15	the beginning of
15	11	the end of the
15	15	the opposite sex
15	14	there has been
15	15	to deal with
15	11	up to the
15	16	would be a
14	13	an example of
14	13	it can be+V3
14	15	do not want
14	15	in the future
14	15	the issue of
14	12	the question of
14	13	the value of
14	13	to go to
14	15	values and beliefs
14	15	will continue to
14	15	will not be
13	13	all of these
13	13	are trying to
13	12	but it is
13	12	can be used
13	13	that there is
13	13	that this is
13	14	the age of
13	14	the lives of
13	13	the opponents of
13	11	to make the
13	13	would not have
12	9	should not be+adj
12	5	aware of the
12	10	don t have
12	13	i believe that
12	13	i do not
12	12	in today s society
12	13	it is also
12	9	it should be
12	13	not have to
12	10	the cost of
12	14	the people of

12	12	there are some
12	6	there have been
12	6	there would be
12	6	those who are
12	13	to be more
12	11	to do with
12	6	when it comes
12	5	would have to
11	11	a lack of
11	11	a matter of
11	10	as a whole
11	14	be used to
11	10	both sides of
11	10	due to the fact
11	9	if you are
11	13	is not only
11	11	it is important
11	12	not be able to
11	11	over the years
11	10	the concept of
11	10	the consequences of
11	13	the first amendment
11	6	would like to
10	9	a great deal of
10	11	a little more
10	11	a strong argument
10	9	and it is
10	12	at the same time
10	9	changed people s lives
10	11	crime does not pay
10	8	do not want to
10	9	has become a
10	11	has been a
10	11	if they are
10	11	in many cases
10	11	in our society
10	10	in some cases
10	12	in terms of
10	11	in the public
10	7	in this case
10	10	many of these
10	9	more and more
10	7	people who are
10	7	that there are

10	10	that they have
10	12	that they were
10	10	that they will
10	11	the life of
10	11	the loss of
10	10	the problem is
10	11	the problem of
10	11	the supreme court
10	12	what they are
10	8	would have been
10	11	you can t
9	10	and that is
9	9	and this is
9	11	argue that the
9	11	at the end of
9	10	based on the
9	10	because of this
9	9	first of all
9	8	going to be
9	11	have come to
9	10	i think that
9	9	if a person
9	8	in front of
9	11	in my opinion
9	11	in the early
9	11	in which the
9	6	is going to
9	7	it comes to
9	7	large number of
9	9	no matter what
9	9	one of these
9	7	should not have
9	8	so that they
9	5	some of these
9	7	that can be
9	6	the general public
9	10	the idea that
9	7	the power to
9	10	the rest of the
9	8	there will be
9	9	they have to
9	8	they should be
9	10	to go through
9	10	to have an

9	9	to prove that
9	9	to stay home
9	10	to take the
9	10	when it comes to
8	7	around the world
8	7	as it is
8	8	at the beginning of
8	7	at the time
8	10	can be seen
8	9	capital punishment is
8	8	does not have
8	8	does not mean
8	8	example of this is
8	7	have the right to
8	8	have to do
8	9	have to go
8	8	he does not
8	8	i have seen
8	6	i m not
8	9	in addition to
8	8	in our lives
8	9	in the article
8	9	in the morning
8	10	in this country
8	8	involved in the
8	5	it is only
8	9	people do not
8	10	people feel that
8	9	should be allowed to
8	8	that could be
8	9	that do not
8	9	that is the
8	10	that it has
8	9	that they can
8	7	that they would
8	9	the american people
8	10	the child is
8	7	the majority of the
8	9	the people that
8	10	the problem that
8	8	the proponents of
8	6	the same as
8	9	the types of
8	9	they will be

8	9	to be in
8	9	to believe that
8	7	to get the
8	7	to have sex
8	8	to keep the
8	9	to think that
8	9	to try to
8	8	what is the
8	8	when i was
8	8	when they are
8	7	will be able to
8	7	will have to
8	7	would be able to
7	8	a form of
7	5	a good example
7	9	a member of
7	9	a number of
7	7	a sense of
7	8	a way of
7	8	along with the
7	8	amount of money
7	8	an attempt to
7	9	an increase in
7	9	are forced to
7	8	are willing to
7	6	being able to
7	9	believe that they
7	8	church and state
7	7	considered to be
7	5	i have been
7	8	i think it
7	7	i would like to
7	7	if there is
7	8	in a way that
7	8	in high school
7	7	in many ways
7	5	in the classroom
7	7	in the long run
7	5	in the media
7	9	in which they
7	7	is a major
7	9	is because of
7	7	is hard to
7	8	is not the only

7	7	is not to
7	7	is the root of
7	8	is very important
7	5	it is obvious that
7	9	it is possible
7	5	more important than
7	9	more time to
7	7	most of the time
7	6	not be allowed
7	6	on their own
7	5	one can see
7	7	people s lives is
7	7	rules and regulation
7	9	should be able to
7	7	should be the
7	7	such as these
7	5	take care of
7	8	that is not
7	5	that they may
7	7	the ability to
7	5	the best way to
7	8	the conflict theory
7	8	the discovery of
7	5	the history of
7	6	the minds of
7	5	the presence of
7	7	the results of
7	8	the united states goverment
7	5	the united states is
7	8	the work place
7	6	there are two
7	8	they are more
7	8	they can be
7	7	they feel that
7	8	this argument is
7	8	this can be
7	7	this is a very
7	8	to live in
7	7	to look at
7	6	to pay for
7	6	to support the
7	8	we don t
7	7	we need to
7	7	who have been

APPENDIX C. SAMPLE LESSON PLAN

Subject: Giving reasons, identifying causes and effects in writing (Classroom Activity Materials - Sample 5)

Grade: Tertiary Level **Time:** 90 minutes

Learning objectives: By the end of this lesson, students will be able to:

- access new phrases to give reasons, identify causes and effects on their own.
- use the COCA to find out about new words and phrases related to a specific topic.
- use phrases in appropriate co- and contexts.
- use different options corpus tools provide.
- use known and new lexical phrases in their writings.

	Time	**Implementation**	**Materials**
Warm-up	10 min.	Introduction to the lexical phrases and their uses. Show examples and usages with english-corpora.org.	Computers/ Projector
Activity 1 (Tasks 1-3)	15-20 min.	a) Divide students into pairs or small groups. b) Give each group a worksheet prepared in advance by the teacher. c) Have students do the tasks from 1-3. d) After a few minutes, have each group share their findings with the class.	Computers/copies of the worksheets
Activity 2 (Tasks 4-6)	15-20 min.	Students complete the tasks 4-6 by consulting the COCA.	Computers, worksheets
Activity 3 (Tasks 7-10)	15-20 min	Students complete the tasks 7-10 by consulting the COCA.	Computers, worksheets
Discussion	10 min.	a) Students present their findings to the exercises. b) Comparison and correction of the students' results.	Projector/ Smart Board
Review	10 min.	Students answer the review questions.	Review questions
Post-task/ Homework/ Extension	5 min.	a) Have students research the use of lexical bundles in different academic disciplines. b) Assign an essay to be written in their choices of interests by using as many phrases as possible.	Computers

*Your students can also use their mobile phones instead of computers.

**Teacher/Learners may also prefer using any other corpus tools they are most familiar with.

***Subjects may change according to the needs of the students.

Worksheet for Activities

Task 1 Search COCA or other academic corpora, or KWIC search in BNC-English corpora (sorting to the left) to find recurrent patterns of use of the noun ***reason***. Look on both sides of the keyword ***reason***, focusing specifically on the phrases in which it is used. Write down three sentences using those phrases.

Task 2 Some frequent patterns from the Academic General corpus are given below. Search academic corpora for some examples illustrating these patterns, and write your own examples in your writing assignment.

the reason to **+ verb**

the reason to believe/suppose/think **+ that + clause**

There is good reason to believe, therefore, that this one sample is probably representative for the species, at least in this area.

the reason that **+ clause**

……

……

……

the reason why **+ clause**

……

……

……

reason for **+ Ving**

reason for **+ noun**

……

……

……

for this/that reason **+ clause**

for the simple reason that **+ clause**

……

……

……

***by reason of* + noun**

……

……

……

Task 3 a) Explore adjectives modifying the noun ***reason*** in academic writing by search-ing COCA, or KWIC search in BNC-English corpora or SkELL (sorting to the left) and look for adjectives:

logical, fundamental……

b) Using all three corpus tools find strong and tentative adjectives used to describe reasons:

strong *valid, apparent*

……

tentative *plausible, unclear*

……

Task 4 Run a search in BAWE on the noun ***reason*** with associated word *no* within three words on the left. Compare three examples from BAWE:

(a) to a different way of expressing the same idea

(b). Do the authors give reasons here? What do you think is the function of the negation?

a) 1…. As there is ***no reason*** to doubt the validity of these results, they will be retained for the analysis.

2. …there is ***no reason*** to assume that there can be one homogenous explanation…

3. …there is ***no reason*** that we should compare these two planets.

b) 1. *These results are valid and will be retained for the analysis.*

2. *There is no single homogeneous explanation.*

3. *We should not compare these two planets.*

Why do writers in the corpus use the phrase ***no reason to*** (a) rather than (b)?

Task 5 a) Consult the COCA to find instances where the noun ***reason*** co-occurs with modal verbs on the left. What search can be done for that? Some concordance lines are presented below. Can you explain why ***reason*** co-occurs with modal verbs?

> measured were not accounted for and this could be a REASON for the deviation from the expected N4 reading for the ending points. This could be another REASON for getting not accurate results. Conclus time climbing than they do walking. There must be another REASON why the characteristics have remained and uilt voluntary start of the process may prove to be another REASON for the new legal rights not to have a ople realise depression can be treated and this may be one REASON the figures for depression have risen. n courts". He advocates that consideration should only be a REASON for a promise rather than vital to

b) Which modal verb(s) most frequently collocate(s) with ***reason***?

can

could

may

might

should

must

would

Try to rank these verbs in order of frequency, then run a search in BNC_English corpora to check. In Collocates, put ***reason*** in the search box, and in the **Collocates** box choose **verb.MODAL** (Figure 2).

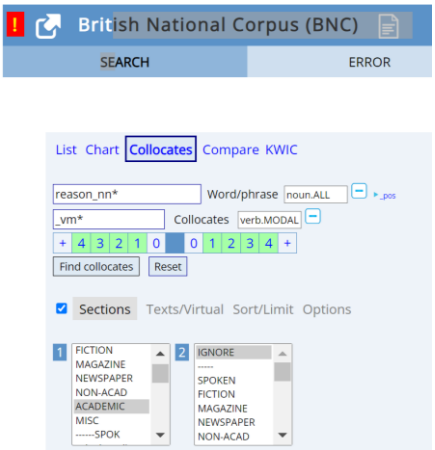


Figure 2 Modal verbs + *reason*. Collocates search in BNC-English corpora.

Task 6 Run concordance searches on other ways of giving reasons (below), and note the sentence patterns in which they are used (i.e. what these phrases are followed by in the corpus.) Is it a noun, a gerund (an ***-ing*** form that functions as a noun) or a clause? For example, ***because*** is consistently followed by a clause, and ***because of*** is usually followed by a noun (Table 2).

Table 2 Giving reasons: sentence patterns

Word or phrase	*Followed by … + example*
because	*+ clause* *This is important because the differences relate to identity.*
because of	*+ noun or noun phrase* *River basin systems will have differences, because of their different environments*
is (may be) due to	
as a result	
results in/ from	
is caused by	

Task 7 Fill in the blanks with the correct phrases given below. There are several possibilities. Use a corpus to check your answers if you are unsure what these phrases are followed by.

result in, result from, because, because of, is /may be due to, as a result of, is caused by, leads to, cause, the reason for, the cause of

1. ……*the fall in the prices, consumers are made better off by vertical integration*.

2. *Population can change* ……*comparative changes of birth rate and death rate.*

3. *It was estimated that closure would* ……*a net saving of roughly £2,923 per annum*.

4. *The waves on the surface of a pond are* …… *dropping a pebble into it.*

5. *The radiation would* ……*an orbiting electron losing energy and collapsing into the nucleus*.

6. *Illness* …… *loss of appetite*

Task 8 Use different ways of expressing reason, cause, and effect to paraphrase the sentence that follows.

> *A reduction in interest rates will result in a fall in the size of this subsidy*….
>
> *(Based on Lextutor, Academic General)*

because, because of, be due to, as a result, results in, result from, result of, to be caused by, lead to, a cause of, to cause, can be attributed to, may be connected with

Task 9 Complete the sentences below using expressions of cause and effect from Task 8. Give several possible causes for each.

1. *Some students experience difficulties with academic writing …*

2. *Severe weather patterns …*

3. *An ageing population …*

Task 10 Run a search on ***due to***, and sort on the left of the keyword. Consider the following examples from concordance lines (BAWE, Lextutor) with *due to*; what hedging devices make a statement more tentative in these sentences? What is the reason for their use?

1. This difference can be ***due to*** *multiple factors*

2. Pain in the hip area may be ***due to*** *pathology in the lumbar spine*

3. I think this is ***due to*** *better information networks*

*4. The increased loss from grazing is predominately **due to** excretion from livestock*

*5. Conflict arises generally **due to** the natural tendency of each member to maximize their profit margin.*

Step-by-step Guideline

Fig. 1: Selecting a corpus.

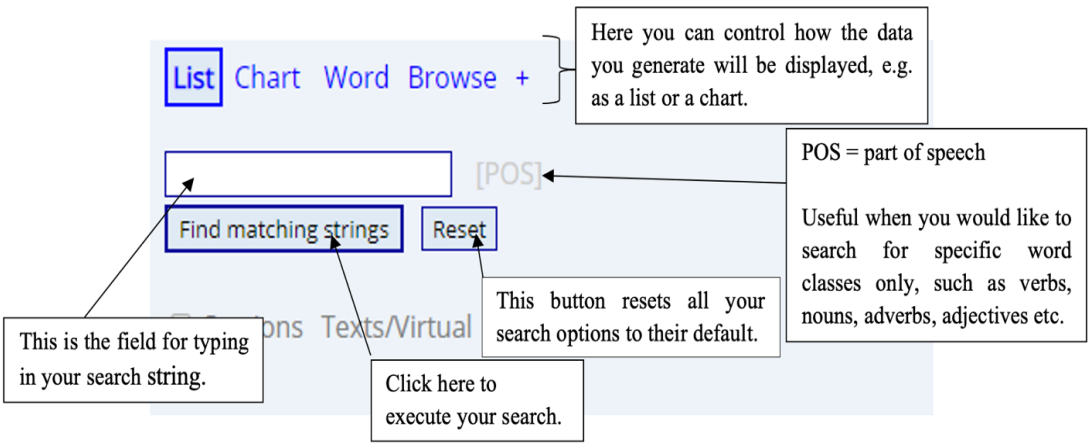


Fig. 2: Conducting a search in the COCA.

Fig. 3: Conducting a search in the COCA.

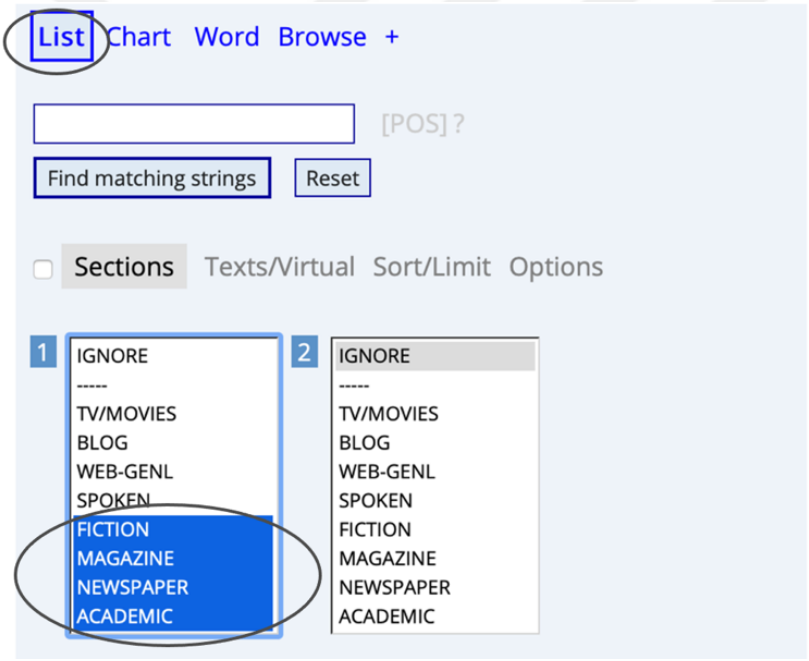


Fig. 4: Limiting the search to specific registers.

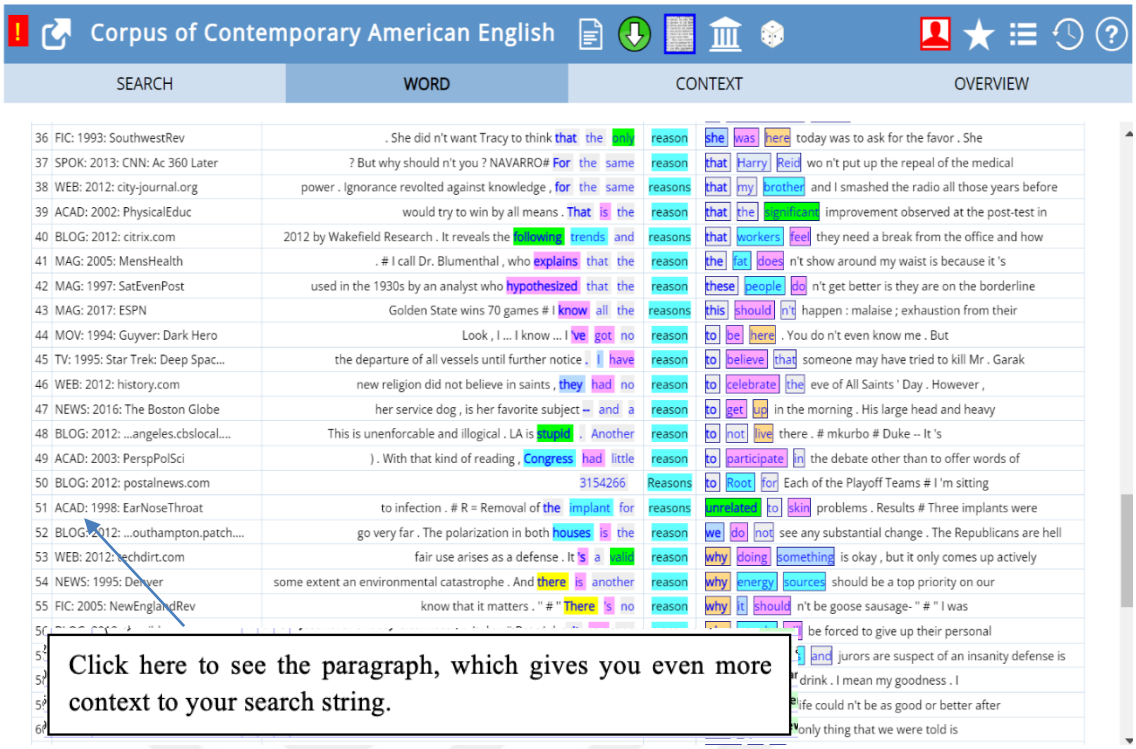
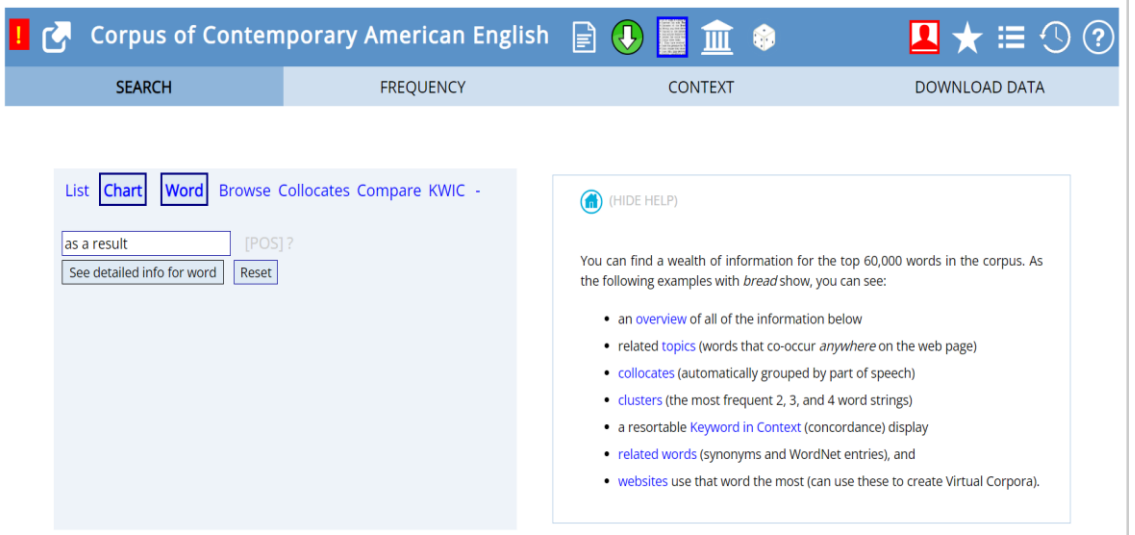


Fig. 5: Result page from the corpus query.



.

Fig. 6: Search page for searching the phrases.

Review Questions:

1. Was using the COCA helpful for learning more about word/phrase/grammar usage in English?

2. Do you think it is worth spending the time learning how to use the different functions corpus tools provide? Why/why not?

3. Are you likely to use the COCA on your own in the future? Why/why not?

APPENDIX D. CLASSROOM ACTIVITY MATERIALS

SAMPLE 1: LEXICO-GRAMMATICAL PATTERNS

Task 1 Have a look at the following sentence:

Avoid to make *a similar mistake in your reasoning.*

Run a search in English-Corpora.org (COCA) or Lextutor Concordancer on ***avoid***, look on the right of it and decide whether the sentence is correct.

Task 2 a) Search academic corpora in Lextutor, choosing ***recommend***, ***suggest***, ***plan***, ***expect*** as keywords; sort to the right; and find out what these verbs are followed by. Take note of your results in Table 1.

Table 1. Lexico-grammatical patterns of *recommend, argue, plan, expect*

Word	*Followed by*				
	Noun	*Noun with preposition*	*Gerund*	*Infinitive*	*Clause*
recommend					
argue					
plan					
expect					

b) Now determine which of these sentences are correct based on your findings from the corpus:

1. It is ***recommended*** *that this method is used throughout.*
2. *I* ***suggest*** *to use this method.*
3. *It was* ***planned*** *to resume the lab work in June.*
4. *We can realistically* ***expect*** *finishing this project in September.*

Task 3 Consider the following sentence and decide whether this usage is correct:

This is the reason because this method was used…

Check by searching the word ***reason*** in at least two corpora and find out what words or structures it is usually followed by. Correct the sentence.

Task 4 Consider the following sentences:

1. *The team of researchers* ***discussed that*** *this method is important.*
2. *The team of researchers* ***discussed why*** *this method is important*
3. *The team of researchers* ***discussed about introducing*** *this method.*
4. *The team of researchers* ***discussed how to introduce*** *this method.*
5. *The team of researchers* ***had a long discussion about*** *introducing this method.*

Are all sentences correct? Check the usage of the verb ***discuss*** in the academic corpora of Lextutor Concordancer. What are the verb ***discuss*** and the noun ***discussion*** followed by?

Task 5 Consider the following sentences and determine which of these sentences are correct. Compare usage of ***because*** and ***because of*** in the concordance lines below. Is there a difference in sentence structure? (Figure 1)

Such errors occur because of a faulty computer control system.
Such errors occur because of computer control system was faulty.

Figure 1. *Because* and *because of* in BAWE corpus in Lextutor.

Can you formulate a rule related to their usages?

Because of is followed by

Because is followed by

The rule is:

Task 6 Consider the following sentence and determine whether the sentence is correct:

This method has its flaws, such as it can only be used under controlled conditions.

Now, check by using the concordancer tool. Search for ***such as***, sort on the right. What is ***such as*** usually followed by? Can you formulate a rule of usage of ***such as*****?**

Task 7 Compare concordance lines containing ***for example*** and ***such as.*** What is the difference in usage of ***for example*** and ***such as*** in terms of the pattern that typically follows these phrases?

such as is followed by ……

for example is followed by ……

The rule is: ……

Adapted from: Karpenko-Seccombe, T. (2021). Academic Writing with Corpora: A Resource Book for Data-Driven Learning. London and New York: Routledge, Taylor and Francis Group.

SAMPLE 2: LEXICAL PHRASES

Task 1. Here are some phrases that students often get wrong. Which form is correct?

a) *The structure of the first chapter is* ***as following / following / as follows / as follow***

Try **equals** search sorting to the left, and use ***follow***, ***follows*** and ***following*** as keywords to find out which phrases are correct.

b) ***As regard / regards / as regarding / with regards / as regards/ with regard to*** *housing, employers' policies vary.*

Task 2. Similarly run a search on ***regard***, ***regards*** and ***regarding***, and sort to the left to find out which phrases are correct in Academic General or any other academic corpus in Lextutor. Take note of your results.

Adapted from: Karpenko-Seccombe, T. (2021). Academic Writing with Corpora: A Resource Book for Data-Driven Learning. London and New York: Routledge, Taylor and Francis Group.

SAMPLE 3. MAKING CONNECTIONS WITHIN AND BETWEEN SENTENCES

Connecting parts of sentences – clauses

Task 1 Below is a screenshot from the English-Corpora.org (COCA) search for synonyms of the linking word ***so***. Think of a sentence with 'so', and write it down. And then, replace 'so' with each of these synonyms. Ask yourself, do all of them replace 'so' accurately, preserving the meaning of the sentence? (Figure 1)

Figure 1. English-Corpora.org (COCA): *so.*

Consider three sentences from the BAWE corpus and work out the differences in meaning between ***consequently***, ***subsequently*** and ***accordingly***. Are they interchangeable?

> *The British left were divided and weak in the 1930s;* ***consequently*** *they were unsuccessful in creating a lasting welfare settlement.*
>
> *Neanderthals are the direct descendants of these first 'true humans' and* ***subsequently*** *we are the direct descendants of Neanderthals.*
>
> *The joint that is hardest struck is usually the hip, and* ***accordingly*** *research has been more focussed in this area.*

Task 2 Fill in the blanks with the appropriate words – ***consequently***, ***subsequently*** or ***accordingly*** –. Even though these words are often given as synonyms, there is one that fits each sentence best.

1. *Such devices can be trained to output desired signals when fed with specific inputs. ……, neural networks can act as controllers in monitoring smart systems.*

2. *A close-ended questionnaire, consisting of three parts, was constructed for a pilot study with 299 respondents and, ……, , for the main study with 675 patrons.*

3. *Proper endoplasmic reticulum (ER) function is crucial to many aspects of cell physiology. ……, , cells react rapidly to correct ER dysfunction.*

4. *…documentation is usually updated ……,.*

B. Register

Task 3 Consider the linking words ***so***, ***thus*** and ***hence.*** Decide which one is more common in formal, academic language, and which is more common in spoken language. And then, run a search on ***so***, ***thus*** and ***hence*** in the BNC-English corpora corpus using the **Chart** tool, which allows us to compare the frequencies of these linking words in texts of different genres. Compare search results. What can you conclude?

C. Usage

Task 4 Concordance lines can also help in understanding the usage of linking words specifically their position in a sentence and associated punctuation. Compare the usage and punctuation of ***consequently*** and ***so.*** Look for recurrent patterns (Figure 2):

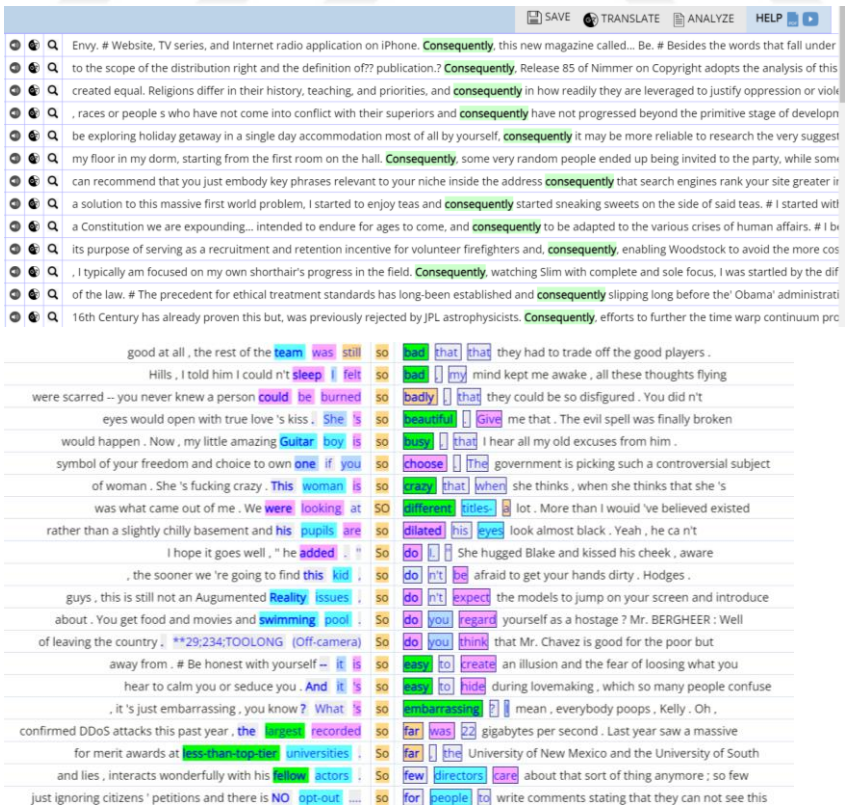


Figure 2. *Consequently* and *so:* usage and punctuation.

Connecting sentences

Task 6 We can use concordances to explore how sentences are connected in academic writing. Study the beginnings of academic sentences in the BNC-English corpora. First choose the KWIC option; in **Sections** choose **Academic**, and search for the beginning of sentences typing a full stop (.) in the search box. Choose three words on the right – this will give you an opportunity to look at how sentences commonly start (Figure 3).

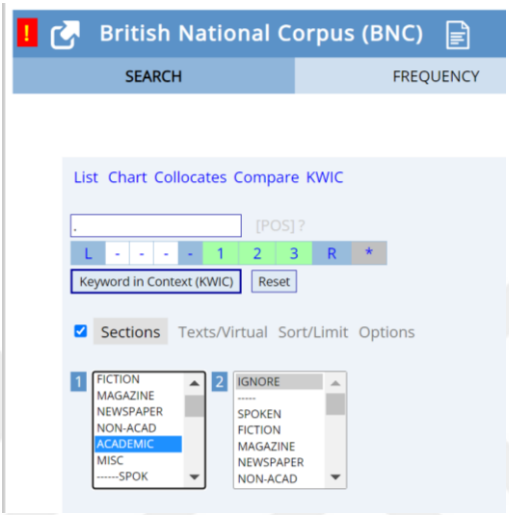


Figure 3. Beginning of sentences search in BNC-English corpora.

Figure 4. Beginning of sentences in BNC-English corpora: search results.

Look at the search results, and do the tasks below:

a) Find the phrases used in academic writing to show connections to the previous or next point.

b) Some of the connecting phrases which provide transitions between sentences:

a. *By sublimation I mean…*
b. *Clearly, however,*
c. *Durkheim was particularly concerned with*
d. *Finally*
e. *For example,*
f. *Further work*
g. *Gifts in kind on a larger scale*
h. *He*
i. *He… also*
j. *However,*
k. *If anything*
l. *In fact*
m. *In other words*
n. *In principle*
o. *In response*
p. *In the case of more pragmatically oriented studies*

Specify which transitions provide connections by

1. offering further explanation

……

2. making clear that some information is added

……

3. narrowing down or generalizing what was said

……

4. referring to what was said before by continuing a thread

……

5. saying something different or opposite to what was said before

……

6. using pronouns to refer to the previous sentence

……

Adapted from: Karpenko-Seccombe, T. (2021). Academic Writing with Corpora: A Resource Book for Data-Driven Learning. London and New York: Routledge, Taylor and Francis Group.

SAMPLE 4: EXPRESSION OF STANCE AND VOICE

Hedges and Boosters

Task 1 Browse several corpora to find examples of hedging devices and write them in the table.

Table 1 . Hedging expressions

Hedges	*Examples*
Modal verbs *may, might, can, could, should, would*	
Verbs *to seem, to appear, to believe, to assume, to suggest, to estimate, to tend*	
Nouns *suggestion, possibility, probability, likelihood, tendency, assumption, estimate*	
Adjectives	
possible, probable, un/likely	
Adverbs • of possibility *perhaps, probably presumably, apparently, admittedly* • of frequency *sometimes, often, commonly, predominantly, occasionally*	
Phrases	

it is believed, to the best of our knowledge, one of, as far as we know, this area is under researched, in some/many cases	

Task 2 Try to rewrite the following sentence with different types of hedging expressions using the hedging expressions from Table 1.

Methods of extraction in the laboratory ***may vary***.

(BAWE, Lextutor)

For example, a stronger way of expressing the same idea would be *Methods of extraction in the laboratory* ***vary***, while the more tentative way would be ***might sometimes vary***. Rank your sentences according to degrees of certainty.

Task 3 Which modal verbs are more frequent than others in academic writing? Rank the modal verbs ***may, might, can*** and ***could*** according to frequency, and then check your result by running a search in BNC-English corpora. Choose the **List** tool and select **verb.MODAL** in the parts of speech (POS) drop-down menu (Figure 1).

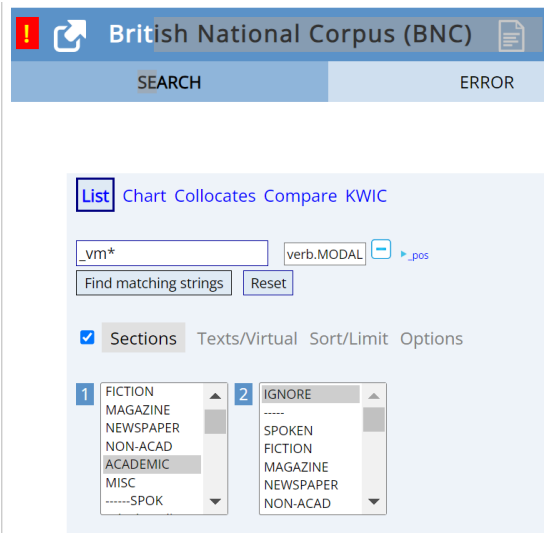


Figure 1. BNC-English corpora List search: *modal verbs.*

Why is it important to have the knowledge of word frequencies?

Task 4 ***Can*** is one of the most common hedging devices expressing possibility:

Better results can be anticipated for the following years.

However, it can also express ability or capability:

Two main results can be derived from the expressions above.

Look through ten results of the following search in Lextutor Academic General, and decide in which cases ***can*** is used as a hedge to express (Figure 2):

A. possibility

B. ability or capability

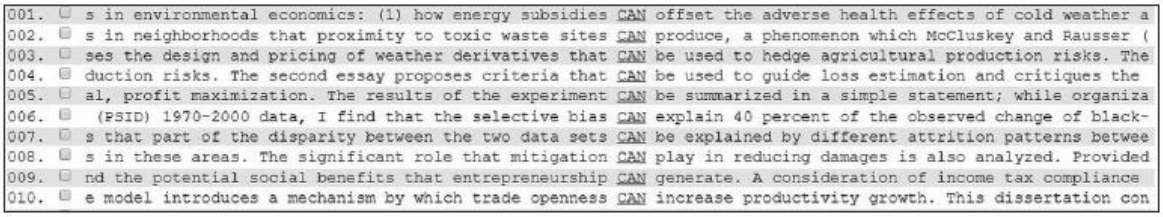


Figure 2. *can* in Lextutor.

Task 5 The following tasks will help you with expressing possibility. Run a family search on *possible* in several corpora to see which forms are used most frequently in academic writing to express possibility. Knowing which forms are most common will allow you to make decisions on which sentence structure to choose and will help with paraphrasing too. Here are the frequencies from BAWE:

possible = 3,639 *possibility* = 765 *possibly* = 672 *impossible* = 576

possibilities = 204 *impossibility* = 39 *impossibly* = 3

Alternatively, use BNC-English corpora wild card search in **List**, putting ****pos-sib**** in the search box. You will then see the list of all word forms in the corpus and their frequencies. Search ****possib**** in KWIC to see concordance lines with various derivatives of ***possible***.

a) Paraphrase the following sentences (abridged from BAWE in Lextutor) using the searches you carried out earlier:

Thus there is no real possibility of bias.

There is also a possibility of physicians making mistakes.

b) Write three sentences related to your essay or research topic using various forms of *possible* and sentence patterns from the earlier searches, e.g.:

There is a real possibility that/of…

Another possibility in relation to…

…does not preclude the possibility of…

This significantly reduces the possibility of…

Task 6 Search for *possibility* in BAWE, Academic General or any specialized BNC corpus, and sort to the left.

a) You will notice several strong patterns of usage of *possibility: the possibility of*, *the very possibility of* and *one possibility* + verb. Which one describes a stronger or weaker possibility?

b) What adjectives commonly modify *possibility*? Take note of the ones that make *possibility* stronger (boosters) or weaker (hedges). Write down your results:

stronger

……

weaker

……

Task 7 Investigate the use of ***may*** in the following MICUSP Critique/Evaluation papers in engineering (CEE and IOE) below. What does ***may*** express in the following cases?

A. a possibility of something happening in the future

B. a tentative suggestion about something

Glacier Way Dam Site Soils Evaluation	Civil & Environmental Engineering	Research Paper
1. The soils have been found to be clayey sand and poorly graded sand for the fine and coarse samples, respectively. The gradation of the samples meets the Army Corps of Engineers criteria for soil restraint but fails to meet allowable filter permeabilities. Therefore, either a different filter media may be evaluated or design 2 may be chosen. For both soils, the Hazen approximation for permeability is not nearly within reason to allow this shorthand estimation to be used in the field. Unacceptable design 1 has an estimated seepage rate of 32.7 ft3/day compared to an improved 19.6 ft3/day for design 2. If this is deemed excessive several recommendations have been included that will get your project underway with minimal effort. All other data may be found on the data and calculation summary page, Appendix A.		
Gradation Analyses and Permeability Tests for Glacier Way dam project	Civil & Environmental Engineering	Research Paper
1. The seepage loss under your dam may be the single most important parameter in determining the stability of your dam because if the loss is too great, the soil beneath the dam may wash away, destroying the dam with it.		
Characteristics of Concrete Aggregates	Civil & Environmental Engineering	Research Paper
1. The values of absorption and specific gravity for aggregate that is not oven dried before it is soaked in water can be significantly higher that aggregate that is oven dried. (1) Larger particles, especially those over 75 mm, may be too thick to allow water to penetrate the pores to the center of the aggregate particle. Therefore it is critical that the procedure state if the aggregate was oven dried and, if so, how long it was submerged in the water. In this case, since the data is being found for a concrete mixture (where the aggregate will be in its natural, moist state) the aggregate was continuously submerged.		
Molding Temperature Analysis for Liquid Rubber Supplier	Industrial & Operations Engineering	Critique/Evaluation
1. As shown above, the last observed data point in the sample makes the individual and moving range control charts for CIR unstable. This sample reveals that there may be a special cause of variation present due to the spike created from one of the last data points in the sample.		

Figure 3. *may* in MICUSP.

And now, compare these results to the argumentative essays of student papers in English. Discuss the similarities and differences.

Adapted from: Karpenko-Seccombe, T. (2021). Academic Writing with Corpora: A Resource Book for Data-Driven Learning. London and New York: Routledge, Taylor and Francis Group.